Dynamics and Structure in Competitive Social Systems

by

Sears Merritt

B.S., University of Colorado, 2004

M.S., University of Colorado, 2010

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Computer Science

2013

This thesis entitled:
Dynamics and Structure in Competitive Social Systems
written by Sears Merritt
has been approved for the Department of Computer Science

_____

Aaron Clauset

_____

Prof. James Martin

_____

Prof. Michael Mozer

Date _____

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

Merritt, Sears (Ph.D., Computer Science)

Dynamics and Structure in Competitive Social Systems

Thesis directed by Prof. Aaron Clauset

Competition is ubiquitous in complex social systems, from informal online gaming environments to professional sports, to workers competing for jobs in the labor market. Given its universality, understanding the complex relationships between a competition's dynamics, competitor behavior, and environmental structure is of great interest to the scientific community and society at large. In this thesis, we analyze the structure of and behavioral dynamics in three different competitive social systems: a massive online game, four professional sports, and a network of occupations. Our results shed new light on how a competition's environment can predict its dynamics, what types of temporal and pro-social behaviors are indicative of friendship between competitors, how those friendships evolve over time, and how the structural properties of an occupation network inform career path decisions and forecast the long term distribution of incomes in an economy.

Dedication

This dissertation is dedicated to Shaina, Ellis and Brooks.

Acknowledgements

This thesis would not have been possible without the guidance, wisdom and support of Professor Aaron Clauset. I would also like to acknowledge the support and insightful contributions of Abigail Jacobs, Christopher Aicher, and the rest of the Clauset Lab, the Computer Science Department, and of course, the University of Colorado.

Contents

Chapter

Tables

Table

Figures

Figure

Chapter 1

Introduction

Prior to the information age, social scientists were limited to studying individual and group level human behavior through manual data collection efforts and small controlled experiments [37, 34]. Since that time, inexpensive computational power and communications infrastructure has developed, matured, and embedded itself into nearly every aspect of people's lives. This trend has allowed society to generate massive swaths of detailed data that characterize complex individual and group social behaviors [70]. To take advantage of this data, computational social science has emerged as a scientific discipline that aims to study social phenomena using these data in combination with quantitative methods drawn from a variety of domains including probability, statistics, machine learning, and network science. With these tools in hand, the field has validated existing theories as well as invalidated others [9, 45].

Competition is ubiquitous in complex social systems, from informal online environments to professional sports, to interactions between job seekers and employers. Given its universality, understanding the complex relationships between a competition's dynamics, competitor behavior, and environmental structure continues to be of great interest to the scientific community and society at large. In this thesis we study three different competitive social systems and shed new light on the structural properties of an occupation network, how a competition's environment influences it dynamics, and what types of periodic behaviors between competitors are indicative of friendship.

1.1 Contributions

In this thesis we make five primary contributions in the forms of statistical analysis, predictive modeling, and complex networks. These contributions are:

Contribution 1: Predictive model of friendship

In many complex social systems, the timing and frequency of interactions between individuals are observable but friendship ties are hidden. Recovering these hidden ties, particularly for casual users who are relatively less active, would enable a wide variety of friendship-aware applications in domains where labeled data are often unavailable, including online advertising and national security. Here, we investigate the accuracy of multiple statistical features, based either purely on temporal interaction patterns or on the cooperative nature of the interactions, for automatically extracting latent social ties. Using self-reported friendship and non-friendship labels derived from an anonymous online survey, we learn highly accurate predictors for recovering hidden friendships within a massive online data set encompassing 18 billion interactions among 17 million individuals of the popular online game *Halo: Reach*. We find that the accuracy of many features improves as more data accumulates, and cooperative features are generally reliable. However, periodicities in interaction time series are sufficient to correctly classify 95% of ties, even for casual users. These results clarify the nature of friendship in online social environments and suggest new opportunities and new privacy concerns for friendship-aware applications that do not require the disclosure of private friendship information. This chapter was published as [84].

Contribution 2: Analysis of social network dynamics in an online game

Online multiplayer games are a popular form of social interaction, used by hundreds of millions of individuals. However, little is known about the social networks within these online games, or how they evolve over time. Understanding human social dynamics within massive

online games can shed new light on social interactions in general and inform the development of more engaging systems. Here, we study a novel, large friendship network, inferred from nearly 18 billion social interactions over 44 weeks between 17 million individuals in the popular online game *Halo:Reach*. This network is one of the largest, most detailed temporal interaction networks studied to date, and provides a novel perspective on the dynamics of online friendship networks, as opposed to mere interaction graphs. Initially, this network exhibits strong structural turnover and decays rapidly from a peak size. In the following period, however, both network size and turnover stabilize, producing a dynamic structural equilibrium. In contrast to other studies, we find that the Halo friendship network is non-densifying: both the mean degree and the average pairwise distance are stable, suggesting that densification cannot occur when maintaining friendships is costly. Finally, players with greater long-term engagement exhibit stronger local clustering, suggesting a group-level social engagement process. These results demonstrate the utility of online games for studying social networks, shed new light on empirical temporal graph patterns, and clarify the claims of universality of network densification. This chapter was published as [83].

Contribution 3: Generative model of competition

In most professional sports, playing field structure is kept neutral so that scoring imbalances may be attributed to differences in team skill. It thus remains unknown what impact environmental heterogeneities can have on scoring dynamics or competitive advantages. Applying a novel generative model of scoring dynamics to roughly 10 million team competitions drawn from an online game, we quantify the relationship between the structure within a competition and its scoring dynamics, while controlling the impact of chance. Despite wide structural variations, we observe a common three-phase pattern in the tempo of events. Tempo and balance are highly predictable from a competition's structural features alone and teams exploit environmental heterogeneities for sustained competitive advantage. Surprisingly, the most balanced competitions are associated with specific environmental heterogeneities, not

from equally skilled teams. These results shed new light on the design principles of balanced competition, and illustrate the potential of online game data for investigating social dynamics and competition. This chapter was published as [81].

Contribution 4: Statistical properties of scoring dynamics in sports

Despite growing interest in quantifying and modeling the scoring dynamics within professional sports games, relative little is known about what patterns or principles, if any, cut across different sports. Using a comprehensive data set of scoring events in nearly a dozen consecutive seasons of college and professional (American) football, professional hockey, and professional basketball, we identify several common patterns in scoring dynamics. Across these sports, scoring tempo—when scoring events occur—closely follows a common Poisson process, with a sport-specific rate. Similarly, scoring balance—how often a team wins an event—follows a common Bernoulli process, with a parameter that effectively varies with the size of the lead. Combining these processes within a generative model of gameplay, we find they both reproduce the observed dynamics in all four sports and accurately predict game outcomes. These results demonstrate common dynamical patterns underlying within-game scoring dynamics across professional team sports, and suggest specific mechanisms for driving them. We close with a brief discussion of the implications of our results for several popular hypotheses about sports dynamics. This chapter was published as [82].

Contribution 5: Occupation networks

Using data from nearly 1,000,000 publicly available resumes and 6,000,000 occupation descriptions, we construct a rich network of occupations and study its structural properties. We find that high degree nodes require social and administrative skills. The number of common skills shared by a node and its neighbors decreases with degree. Finally, we identify a region of the network that is vulnerable to an income trap, where the mean salary of a node's neighbors is lower than its own. We conclude with a brief discussion of how these

results can be used to inform career path decisions.

1.2 Roadmap

The remainder of this thesis is organized into seven distinct chapters. The first focuses on relevant quantitative methods and data processing techniques used throughout the research. The following five chapters present preliminary results of our research while the final chapter presents conclusions and topics for future work.

Chapter 2

Background

2.1 Competitive social systems

Competitive social systems are complex, dynamical systems whose properties are constructed by the interactions of many individuals. In many cases these systems contain memory, counterintuitive feedback loops, and non-linear relationships. Given these complexities, how do we study such systems? To analyze dynamics, physicists and mathematicians have applied modern techniques from statistical mechanics and chaos theory to these problems, often times with great success [50, 51]. From a structural point of view, network science has emerged as a suite of tools and methods that can be used to measure structural features and connect these features with the underlying dynamical mechanisms [119, 95].

2.2 Stochastic processes

A stochastic process is a set of random variables used to represent the state of a system, usually as a function of time. $\{X_t : t \in T\}$, where T is the known as the index set, often interpreted as time [118]. When the set of random variables obeys the following relation $\Pr(X_n = x|x_0, x_1, \ldots, x_{n-1}) = \Pr(X_n = x|x_{n-1})$ the system is known as a 1^{st} order Markov process and can be represented as a Markov chain. That is, the probability that the system is in state x_i is dependent only on its previous state, x_{n-1}. When $\Pr(X_n = x)$, which indicates that the system is independent of all prior states including the previous one, the system is a Bernoulli process [118].

2.2.1 Bernoulli processes

A Bernoulli process is often thought of as a coin flipping process where the probability of observing a head is equal to c and the probability of observing a tail is equal to $1 - c$. The probability of observing k heads in n trials can be computed as

$$\Pr(\text{k heads in n trials}) = \binom{n}{k} c^k (1-c)^{n-k}. \tag{2.1}$$

In this case the ordering of the k heads is irrelevant. Similarly, the probability of observing between 0 and m heads in n trials can be computed as

$$\Pr(0 \leq k \leq m \text{ heads in n trials}) = \sum_{k=0}^{m} \binom{n}{i} c^i (1-c)^{n-i}. \tag{2.2}$$

To estimate the parameter c from a sample of size n, the method of maximum likelihood can be used [118]. The likelihood function of the Bernoulli distribution is

$$L = \prod_{i=1}^{n} c^{x_i} (1-c)^{1-x_i} \tag{2.3}$$

where $\{x_i, \ldots, x_n\}$ are the data that take on values of heads or tails. Maximizing this function yields an estimate of c, denoted as $\hat{c}$, equal to $\frac{\sum_{i=1}^{n} x_i}{n}$.

2.2.2 Markov chains

When state changes in a system are not independent, as in the Bernoulli process, Markov processes can be used to represent their dynamics. As mentioned earlier, in the simplest case, where the system is memoryless and its state transitions depend only on the current state, a Markov chain can be used to model its dynamics. Transition probabilities in Markov chains are values which indicate the probability of the system transitioning from state i to state j and are denoted as p_{ij}. When these transition probabilities are time invariant, or homogenous, they can be represented as a matrix $\mathbf{P}$, known as the transition

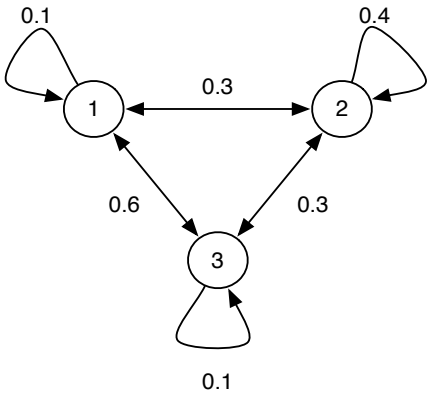

Figure 2.1: Example three state Markov chain

matrix, shown mathematically in Equation 2.4 below and graphically in Figure 2.1.

$$\mathbf{P} = \begin{bmatrix} 0.1 & 0.3 & 0.6 \\ 0.3 & 0.4 & 0.3 \\ 0.6 & 0.3 & 0.1 \end{bmatrix} \tag{2.4}$$

By the Chapman-Kolmogorov equations we can compute the probability of transitioning from state i to state j in n steps, $p_{ij}(n)$, as $\mathbf{P}^n$.

In addition to computing the probability of ending in a state given a starting state, it is also possible to compute the probability of reaching one state before another. Given an nxn transition matrix $\mathbf{P}$, where values in the n^{th} row and column are equal to 1

2.2.3 Poisson process

A Poisson process is a stochastic process used to characterize the rate and number of occurrences of events, such as the number of goals scored in a competition, among others. The most basic assumptions of the process are that events are independent of one another and do not occur simultaneously. In the case where the rate at which events occur, denoted as λ, is stationary, then the Poisson process is homogenous. As a result, the inter-arrival times between consecutive events are distributed according to an Exponential distribution

with parameter $1/\lambda$. Mathematically we can say that for a homogenous Poisson process

$$\Pr(\text{k events in } \tau \text{ intervals}) = \frac{e^{-\lambda\tau}(\lambda\tau)^k}{k!} \tag{2.5}$$

where k is the number of observed events and τ is the number of time intervals. When the process is non-stationary and λ varies with time, then the expected number of events between time i and j is defined as

$$\lambda_{i,j} = \int_i^j \lambda(t)dt \tag{2.6}$$

Thus for the non-homogeneous Poisson process, mathematically we can say

$$\Pr(\text{k events in } \tau \text{ intervals}) = \frac{e^{-\lambda_{i,j}}\lambda_{i,j}^k}{k!}. \tag{2.7}$$

where $\tau = j - i$.

2.3 Complex networks

Networks are a powerful tool for studying complex social systems because they provide a principled way to represent and study the structure produces by dynamical interactions between a system's individual parts. Mathematically, a network is defined as a set of vertices and edges, denoted as $G = (V, E)$ where V is the set of vertices or nodes and E is the set of edges or links [95].

2.3.1 Clustering

The clustering within a network can be used to characterize the connectedness of the network and of individual vertices. At the vertex level, the clustering coefficient measures how close a vertex and its neighbors are to forming a clique. Mathematically, the vertex level clustering coefficient for vertex i is computed as

$$C_i = \frac{\text{number of connected neighbors}}{\text{number of possible connected neighbors}} \tag{2.8}$$

From a sociological point of view, a clique is often viewed as a tightly connected set of friends. Social networks are often times composed of large numbers of sparsely inter-connected cliques. This type of structure gives rise to notion of a small world [119].

2.3.2 Degree distribution

The degree of a vertex is simply the total number of edges that connect it to other vertices in the network. A degree distribution is a distribution indicating the fraction of vertices that possess a particular degree, k. In a directed network, degree distributions can be used to represent edges lead to and from a vertex, known as in-degree and out-degree distributions respectively.

In random networks, the degree distribution is Poisson($\langle k \rangle$), where $\langle k \rangle$ is the network's mean degree. Complex networks, particularly social networks, often times have heavy tailed or power-law tailed degree distributions, expressed mathematically as

$$\Pr(k) = k^{-\alpha} \tag{2.9}$$

where α is known as the scaling exponent. A heavy tailed degree distribution indicates that a large number of vertices have a small number of neighbors while a small number of vertices have a large number of neighbors. When networks exhibit power-law degree distributions[1] , they are known as scale-free networks. This heavy-tailed structure is indicative of an underlying order or hierarchy within the system.

2.3.3 Centrality

Centrality measures are used to indicate a vertex's relative importance. Socially, a vertex's centrality is used to quantify its level of influence on others. From a diffusion point of view, the centrality of a vertex quantifies how fast a vertex can spread information.

[1] See [33] for a rigorous method for verifying the existence of a power law

Closeness centrality: Closeness is computed by taking the inverse of a vertex's farness. Farness is defined as the sum of the lengths of the shortest paths between itself and all other vertices in the network. Thus, when a vertex has a large closeness centrality, it contains many short paths between itself and other vertices in the network.

Betweenness centrality: Betweenness centrality quantifies how often a vertex is a member of a shortest path between pairs of vertices in the network. A vertex with a high betweenness centrality suggests that it acts as a an intermediary in relaying information between vertices in the network.

2.4 Data processing

Processing massive quantities of data requires large scale, parallel processing systems. The de-facto standard tool for this task is Hadoop. Hadoop is a distributed system that implements a distributed file system and the map-reduce framework [53, 42].

2.4.1 Hadoop

Hadoop is a distributed system support by the Apache Software Foundation. It supports distributed data processing applications running on clusters of computers, typically constructed from commodity hardware. Instead of sending data to processing units, the system sends the processing to the data [120].

A Hadoop cluster is composed of a name node, a set of worker nodes, each of which run a task tracker, and a job tracker. The name node maintains the state of the distributed file system by storing a mapping of files to data blocks to worker nodes. The job tracker coordinates and maintains the state of map-reduce jobs by accepting jobs, determining the location of the data through name node queries, and scheduling the work with relevant task trackers.

The system also provides fault tolerance at both the file system and processing levels.

Data is replicated across more than one machine such that the system can sustain node failures and not suffer data loss. At the processing level, if mapper or reducer processes fail, the system can restart, or simply discard the subset of data associated with the failing code.

2.4.2 Map-reduce

The map-reduce framework is designed for the processing of web scale data sets distributed across many machines organized as a cluster. Map-reduce is composed of three distinct phases, a map phase, a shuffle phase, and a reduce phase. In the map phase, data that is structured as key-value pairs is processed in parallel by separate processes running on machines that store the data. The output of the map phase is a set of new key-value pairs. Upon completion of the map phase, the shuffle phase takes the key-value pairs, sorts them and transmits them to a specified number of reducer processes. Each reducer process performs a computation on a set of key-value pairs and then emits a final key-value pair that is written to the underlying distributed file system.

The canonical example of map-reduce consists of counting word frequencies in a corpus. Each mapper accepts a line as input and breaks it into words by splitting the string on whitespace. For each word in each line processed by each mapper, a key-value pair is emitted, where the key is the word and the value is an integer whose value is 1. After the map phase is completed, the key-value pairs are sorted based on the key value. The sorted key-value pairs are partitioned and delivered to the reducers. Each reducer sums the counts for each word and emits a single key-value with the word and sum. While this example is simple, map-reduce has proven to be adaptable and capable of supporting complex algorithms from the domains of machine learning and network science [122, 111].

Chapter 3

Detecting Friendship Within Dynamic Online Interaction Networks

3.1 Introduction

For many online social systems, understanding which users are "friends," can be extremely useful, e.g., for targeted word-of-mouth advertising, product recommendations, or detecting hidden social relationships [1] . In some systems these relationships are provided by the users themselves, but even when the friendships are not explicitly labeled, we can often still observe the timing and character of pairwise social interactions; for example, citations between scientists [41], appearances together in photos [35], exchanges of tweets [121], emails [38] or phone calls, playing games together, purchasing goods or services from businesses, etc.

This raises the question of whether hidden or latent friendship ties can be inferred from such interaction data alone. For most online systems, this is complicated by the typically heavy-tailed distribution in the volume of interactions generated by different users: only a small fraction of users account for the majority of all interactions, providing deep histories from which to learn, while most users are "casual," generating relatively little data. Inferring latent ties from observable interactions promises to create both new opportunities and raise new privacy concerns for friendship-aware applications, e.g., in online advertising, where latent tie inference could facilitate social marketing or better estimate product preferences, and online security, where it could uncover clandestine associations and activities.

[1] This chapter was published as S. Merritt, A. Z. Jacobs, W. Mason, A. Clauset, Detecting friendship in dynamic online interaction networks, International Conference on Weblogs and Social Media, Jul. 2013

For many computational social science questions, online multiplayer games are a rich but underutilized source of detailed, temporal interaction data. Past work in this area has shed light on competitive dynamics, social organization, economic trading networks, and deviant behavior [112, 69, 19]. Here we utilize a massive data set from the popular online multiplayer game *Halo: Reach* to investigate the degree to which latent social ties can be automatically identified from social interaction data alone. This data set contains details on more than 18 billion interactions among more than 17 million unique individuals across 700 million game instances, and serves as a model system by which to investigate the general question of detecting friendship in dynamic online interaction networks.

From these data, we extract a temporal interaction network, in which two individuals are connected at time t if they shared a social interaction at time t. Here, interactions are playing a game together. We annotated each interaction with information about its character and magnitude, e.g., if it was a prosocial or antisocial interaction. We then combine these data with the results of an anonymous online survey of the player population [80], including friendship and non-friendship labels for every individual in their time series.

We then design and study nine statistical features representing temporal and cooperative-type interactions. Temporal features capture interaction patterns via periodicities, interaction volume, and the similarity in actions within the online system. Cooperative features quantify the prosocial character of the interactions such as direct and indirect assistance in scoring points, and "betrayals," the equivalent of scoring on one's own goal in the game, which indicates antisocial behavior toward the betrayed individual. Although our cooperative features rely on in-game data specific to *Reach*, the intention here is to capture the character or sign of the interaction [74], and thus analogous features can likely be constructed for other types of interaction data. For instance, the interaction patterns in the game setting could correspond to check-ins with a location-based application; the cooperative features in the game could correspond to positive or negative comments on an online forum.

From a social theory perspective, temporal features are expected to provide a weaker

signal than cooperative ones because the former ignore the additional information explicitly contained in the latter. On the other hand, temporal features are more generalizable because they can always be derived from interaction time series, even when auxiliary information is unavailable, e.g., to study co-location, online social interaction, and communication data [36, 30, 45, 38]. In contrast to many standard data sets, our data allow us to directly compare the predictive utility of these two types of features.

The self-reported friend and non-friend labels from the online survey allow us to quantitatively measure the accuracy of our latent tie inference methods, and we take a supervised approach to learn which features perform well at this task. We also explore the way their performance degrades as we examine ties with progressively less data, which is an important concern for real-world applications. In general, we find that latent friendship ties can be predicted with over 95% accuracy when two individuals have had at least 10 interactions. This level of accuracy is achievable using either the auto-correlation of interaction (temporal) or the number of assists (cooperation). The total volume of interactions between individuals is also a good predictor, but it is less efficient than our two best features. These results clarify the nature of friendship in online social environments and suggest new opportunities and new privacy concerns for friendship-aware applications that do not require the disclosure of private friendship information.

3.2 Related work

Our work draws from three distinct lines of research. Most uses of online game data have focused on understanding certain aspects of human social behavior in online environments. Examples include individual and team performance [107, 108, 105, 106], expert behavior [65], homophily [63], group formation [64], economic activity [29, 10], and deviant behavior [2]. Most of this work has focused on massively multiplayer online role playing games (MMORPGs), e.g., World of Warcraft, although a few have examined social behavior in first person shooter (FPS) games like *Reach* [105]. Relatively little of this work has

focused on the structure of social networks.

Some studies in social network analysis have considered human behavioral patterns in proximity and periodicity, e.g., questions regarding how the accumulation of interactions over time or physical proximity and geographic location can influence the induced social network structure [30, 45, 38, 35]. Few of these studies have focused on online interactions and the way they reflect underlying social ties.

Another significant thread comes from the literature on link prediction. Several studies have considered the question of predicting links in future time steps based on the pattern of links in the past [77]. Others have focused on predicting hidden or missing links when given a partially observed network [31, 104], and on how similarities in preferences and periodic behavior can predict social ties and their sign (friend or foe, trust or distrust) [1, 76, 45, 35, 74].

Of particular relevance is a recent study that applied a similar approach to ours, with good results, to the more narrow question of distinguishing close and not close friends among a user's ties on Facebook [68]. Otherwise, very few studies have focused on the specific question and context considered here. A distinguishing feature of our study is the use of survey data, which provides us with "ground truth" labels of subjective friendship or non-friendship for observed interactions. By combining these ground-truth labels with the detailed data on pairwise social interactions among all individuals, we directly explore the question of distinguishing mere interactions from genuine latent friendships.

3.3 Data and survey

3.3.1 Game details

Our interaction data are drawn from *Halo: Reach*, a popular online first person shooter game. It was publicly released by Bungie Inc., a former subdivision of Microsoft Game Studios, on 14 September 2010, and has generated more than 1 billion games since. Within

the *Reach* system, individuals choose from among seven game types and numerous subtypes, which are played over more than 33 terrain maps. Games can be played alone or with or against other individuals over the Xbox Live online system, and each individual on the system is identified by a unique "gamertag." Players may choose from among several "playlists," which subdivide the total player population and which are based around specific game types.

Once a playlist is chosen, individuals or small "parties" of players (typically friends) are grouped into teams by an in-game "matchmaking" algorithm. This algorithm is based on the TrueSkill system [60], which attempts to create teams with equal total skill (subject to some practical constraints). When a competition is complete, by default all its players are placed in a new game together, but all players or any subset may choose to reenter the matchmaking process to find new teammates or competitors. Both individual game and individual player summaries were made available through the Halo Reach Stats API.[2]

Through this interface, we collected the first 700 million game instances (roughly 305 days of activity by 17 million individuals). Among other information, each game file includes a Unix timestamp, game type label, and a list of gamertags. Each gamertag is associated with a particular team and a set of attributes indicating specific cooperative behavior actions amongst the individuals, described below. This large database provides us with complete data on the timing and character of interactions between individuals but provides no information about which interactions are produced by friendships versus non-friendships.

3.3.2 Survey

We combine these in-game behavioral data with the results of an anonymous online survey of *Reach* players [80]. In the survey, participants supplied their gamertag from which we generated a list of all other gamertags that had ever appeared in a game with the partici-

[2] The API was active from September 2010 through November 2012. API documentation was taken offline in September 2012.

pant. From this list, the participant identified which individuals were friends. [3] We interpret these subjective friendship labels as ground truth. From these data, we constructed a social network with links pointing from participants to their labeled friends. In our supervised learning analysis, both a labeled friendship and the absence of a label are treated as values to be predicted (i.e., we assume survey respondents explicitly chose not to label their co-player as a friend). Of the 965 participants who had completed the friendship portion of the survey by April 2012, 847 individuals appear in our data (the first 305 days of play); this yielded 14,045 latent friendship ties and 7,159,989 non-friendship ties.

Survey participants were a sparse sample of a large population, and the resulting social network is a composed of mostly disconnected egocentric subgraphs. Labeled friendship ties are directed edges, while observed interactions are bidirectional. We note that because survey participants were recruited through advertising on web fora related to *Halo: Reach*, they are a non-uniform sample of the general *Reach* population, e.g., they tended to be unusually skilled players [80]. Nonetheless, our sample has sufficient variability to demonstrate the general applicability of our results across the player population.

3.3.3 Interaction network

We represent the set of pairwise interactions as an annotated temporal network, in which edges have endpoints, exist at a specific moment in time, and are decorated with auxiliary information on the character and context of the interaction. Vertices in the network correspond to gamertags, and two vertices are connected if they appear in a game instance together at time t (time of day, in 10 minute intervals). Each vertex thus has a sequence or time series of interactions with other vertices. We then annotate each edge with information like whether the corresponding individuals were on the same team, what game type produced the interaction, and number of games played together at time t. The resulting net-

[3] In the survey a friend is defined as a person known by the respondent at least casually, either offline or online.

work, derived from our complete game sample, contains 17,286,270 vertices, 18,305,874,864 temporal edges, and spans 305 days. The subgraph of interactions by our survey participants contained a total of 2,531,479 vertices and 665,401,283 temporal edges over the same period of time.

3.4 Inferring friendship

To recover latent friendship ties given only the time series of annotated interactions between pairs of individuals, we take a supervised learning approach. Using classification trees and a logistic regression classifier [18], we learn which features are best for predicting latent friendship ties. Of particular interest will be computationally lightweight models that could be applied on large scale systems.

The self-reported friendship and non-friendship labels from the anonymous online survey serve as prediction targets. We investigate the accuracy of our statistical features, divided into temporal and cooperative classes and considered individually, for predicting latent ties. Temporal features are derived explicitly from a time series of interactions, without regard to the character or context of those interactions. Cooperative features are derived from the auxiliary data and capture the degree to which an interaction is prosocial. In the construction of several features, we use the massive unlabeled data to derive simple statistical expectations that are used to normalize the raw statistics.

3.4.1 Temporal features

Overall gameplay dynamics within the *Halo:Reach* system are highly periodic (Fig. 4.1), with the peak online population on each day of the week occurring between the hours of 3:00pm and 6:00pm Pacific Standard Time (PST) and the minimum occurring near 4:30am. Since most players reside in the US and the majority of the US population is located on either the East or West coasts, the three hour window of peak play seems likely related to the coasts' three hour time difference. Furthermore, the peak period is roughly synchronized

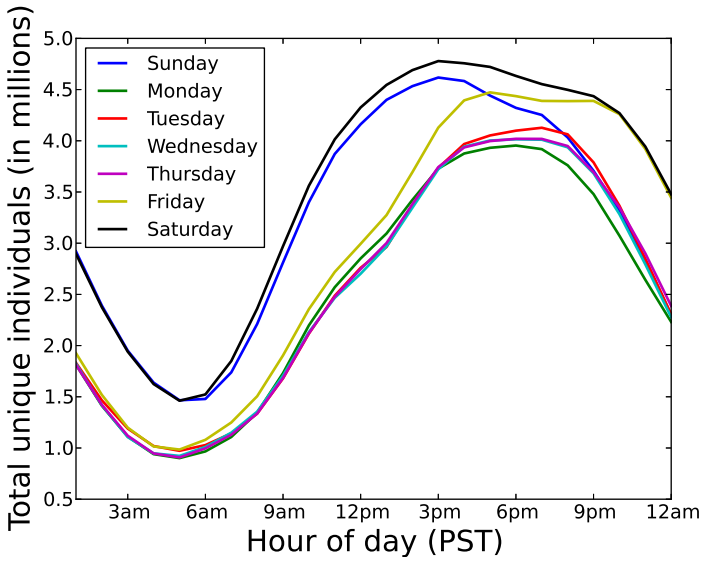

Figure 3.1: Number of unique individuals ever seen at a given time of day (in Pacific Standard Time), across the 305 days spanned by the data, illustrating significant daily and weekly periodicities.

with the class schedules of secondary and post-secondary schools, where the majority of classes occur between the hours of 8:00am and 2:00pm. Finally, we observe a strong weekend effect, with Friday night game play rising to weekend levels, Saturday play remaining high and steady for the majority of the day and night, and Sunday play peaking relatively early and then tapering off after roughly 3:00pm. These regularities suggest several statistical features for capturing latent friendship ties.

Pair autocorrelation. Pairs of individuals in *Reach* that are friends are known to play many more consecutive games (12, on average, or about 2 hours of time) than non-friends (1.25, on average) [80]. Thus, continuous interaction over a significant span of time is likely an indication of a latent tie, while more intermittent interactions likely indicate a non-friend tie, given the large population of non-friends available to play at any time. The expected diurnal and weekly cycles observed in the data will modulate these behaviors, and

a reasonable approach for their quantification is via interaction periodicity. Let

$$n_{x,y}(t) = \mathbb{1}\{x \text{ and } y \text{ play together at time } t\} \tag{3.1}$$

represent the time series of binary interactions between individuals x and y, where 1 indicates an interaction at time t and 0 indicates no interaction. If x and y are friends, we expect $n_{x,y}(t)$ to exhibit stronger periodicity than for non-friends. This expectation may be quantified as the autocorrelation of the time series $n_{x,y}(t)$ over all time lags τ:

$$AC_{x,y} = \sum_{\tau} \sum_{t} n_{x,y}(t) n_{x,y}(t-\tau). \tag{3.2}$$

If $n_{x,y}(t)$ is generated by a non-friend pair, $AC_{x,y}$ should be small because these individuals do not interact regularly. On the other hand, if $n_{x,y}(t)$ is generated by a friend pair, we expect $AC_{x,y}$ to be large.

Pair frequency. A corollary of our previous argument is that friend pairs will likely produce a greater number of interactions over a fixed time period than non-friend pairs. Let N_x be the total number of games played by individual x, and

$$N_{x,y} = \sum_{t} n_{x,y}(t) \tag{3.3}$$

be the number of those games played with individual y. The fraction $N_{x,y}/N_x$ thus captures the share of x's interactions that involve y. Because we expect friend pairs produce more interactions than non-friend pairs, this fraction should be relatively large for a latent friend pair, even if the total number of x's interactions, N_x, is small.

Individual entropy. Recent research has shown that individuals who maintain diverse or unpredictable patterns in their daily schedules in the physical world tend to have larger numbers of friends, as quantified by an entropy measure [36]. But, online environments differ from physical ones in important ways, being more flexible and offering fewer constraints on

"large" movements. It is thus an interesting question whether a digital version of these entropy measures can predict latent social ties as well as its physical analog.

Toward this end, we define entropy measures on an individual's schedule (when they interact), game type (in which game context do they interact), and combined schedule and game type. For a given individual x, we observe the series of x's appearances at "location" $\ell \in \mathcal{L}$, where $\mathcal{L}$ represents the set of all possible locations. We consider three versions of this measure: (i) schedule entropy $H_t(x)$, with locations as days of the week, (ii) spatial entropy $H_s(x)$, with locations as *Reach* "playlists" (which subdivide the full population into groups wanting to play a specific type of game), and (iii) the entropy $H_{s,t}(x)$ over all pairs of schedule and spatial locations.

Mathematically, we compute a given entropy measure as

$$H_{\mathcal{L}}(x) = -\sum_{\ell \in \mathcal{L}} p(x, \ell) \log p(x, \ell), \tag{3.4}$$

where $p(x, \ell)$ corresponds to the observed probability of individual x at location ℓ, i.e., the fraction of all observations of x in which x is observed at location ℓ. We expect the schedule entropy to quantify the diversity of an individual's interactions across time: individuals who typically play on Tuesdays (say, at 8:00pm to meet their friends) will have a lower entropy than those who play in more ad hoc fashions. Similarly, we expect the combined schedule-location entropy to capture regularities such as playing in one game environment on Tuesdays but in different environments over the rest of the week.

For predicting friendships, we take the sum of the individuals' entropies, i.e., $H_t(x) + H_t(y)$, as opposed to a joint entropy measure. A low sum of entropy measures would suggest that both players have low diversity playing patterns, which need not be coordinated. A higher sum would suggest that at least one player of the pair has a more unpredictable schedule; however, knowing this is true for only one player is sufficient to suggest that other temporal signals might be more meaningful. An individual that plays sporadically but with a few regularities (e.g., consistently playing on Saturday mornings with the same

set of individuals) suggests evidence of social coordination. A low entropy pair would then likely be either highly autocorrelated if they played on similar schedules, or exhibit very low autocorrelation if on different schedules. A rich class of temporal features lets us better describe the temporal patterns exhibited by the players in our sample and test existing hypotheses [36].

3.4.2 Cooperative features

Our temporal features explicitly ignore the character of the interactions. Recent work and previous results suggest that friend pairs interact differently than non-friend pairs, and features that capture these differences can be expected to be good predictors of latent ties [58, 80].

Betrayals. One feature of *Reach* that differs from many other online social systems is the ability to commit an explicitly antisocial action, in the form of a "betrayal." These actions are equivalent to an "own goal" and result in a penalty for the betrayer's team. A quirk of the method by which *Reach* places players into a game is that occasionally friends are placed on opposing teams. Past work has shown that when this happens, one team tends to experience an increased betrayal rate as friends on one team turn against their teammates to help their friends on the other team [80].

For a pair of individuals x and y, we capture this tendency by counting betrayals by x that help y, i.e., when x and y are on different teams. Let $b_x(t)$ count the number of betrayals performed by x at time t. Our measure is then

$$B_{x,y} = \sum_t b_x(t)\, \mathbb{1}\{x,\, y \text{ playing on different teams}\}. \tag{3.5}$$

Direct assistance. During a game instance, individuals can provide direct assistance to each other in scoring a point. Like betrayals, this prosocial action can occur with or without deliberate coordination of actions. Because friend pairs are expected to exhibit greater

frequencies of prosocial behavior toward each other, a large number of direct assists should correlate with latent friendship ties.

Let $a_x(t)$ count the number of direct assists performed by individual x at time t. The total number of assists $A_{x,y}$ capture the volume of prosocial behavior on this tie,

$$A_{x,y} = \sum_t a_x(t)\, \mathbb{1}\{x,\, y \text{ playing on same team}\}. \tag{3.6}$$

Indirect assistance. *Reach* also allows an individual to indirectly assist another in scoring points, in which x drives a vehicle while y operates a vehicle-mounted gun. This behavior requires substantially more coordination than direct assists, and thus may provide a more informative measure of latent friendship.

Let $v_x(t)$ count the number of indirect assists attributed to x at time t. The total number of indirect assists from x to y, denoted $V_{x,y}$, is

$$V_{x,y} = \sum_t v_x(t)\, \mathbb{1}\{x,\, y \text{ playing on same team}\}. \tag{3.7}$$

3.4.3 Predicting latent friendships

In our initial exploration of the predictability of latent ties from interaction data, we use classification trees to gain intuition about which features or combinations thereof are likely to be predictive. For this data exploration, the interpretability of classification trees is a strength, compared to, e.g., random forests[4] . Subsequently, we will consider the performance of individual features.

For learning the classification tree, we divided our data into equally sized groups of individuals for testing and training. Cross-validation within the test set was used to control the tree's complexity, pruning branches that did not significantly improve the fit of the model. The resulting tree is highly compact, with only a few features being retained (Fig. 3.2).

[4] To aid interpretation of the tree results, we normalize feature values by the average observed values taken from a uniform random sample of roughly 1 million players. For each of the players in the random sample we compute feature values for each player they interacted with in the data.

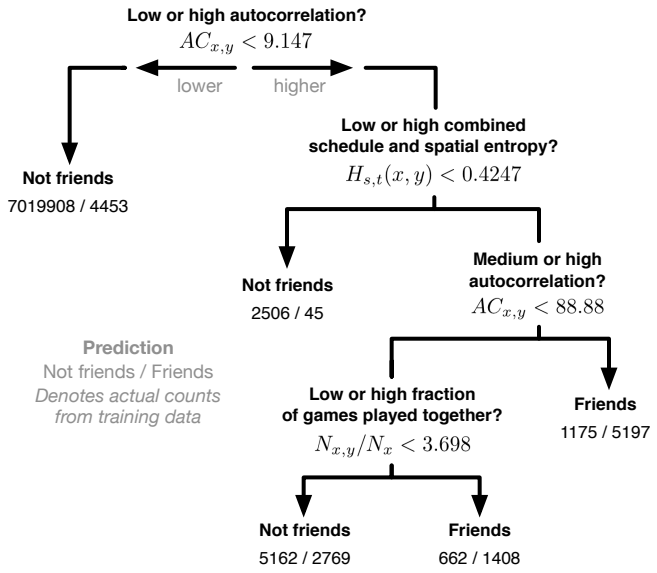

Figure 3.2: A classification tree found using all features except $A_{x,y}$. This tree only uses temporal features, and performs well: the error rate is 0.0013, which is significantly better than the naïve classifier error rate of 0.0020. The out-of-sample AUC for this tree is 0.924. [4]

Repeating our analysis with different subsets of the features and different training and test sets allows us to probe their relative importance and correlation structure.

All of the resulting trees beat the baseline accuracy of a naïve classifier. This baseline is in fact a significant barrier because the number of latent ties is a small fraction (0.2%) of the total number of ties we consider and we can naïvely score well by guessing that every tie is a non-friend. For this reason, we use the Receiver Operating Characteristic (ROC) curve and the Area Under the ROC Curve (AUC) [25], which gives the probability the classifier will rank a randomly selected positive case higher than a randomly selected negative case.

At the level of feature classes, temporal features are most useful for correctly predicting friendship: when trained on all features, the best tree splits first on autocorrelation $AC_{x,y}$, followed by splits on combined schedule and spatial entropy $H_{s,t}(x,y)$, autocorrelation $AC_{x,y}$ (again), and normalized pair frequency $N_{x,y}/N_x$. Similar trees are found when training across all features excluding direct assists $A_{x,y}$, or only temporal features: for all three feature sets (all features, all features except assists, and temporal features only), the final trees yield

	feature	γ	$\hat{\theta}$	$\hat{\sigma}$	$\lvert Z\rvert$	p	AUC
temporal	pair autocorrelation	$AC_{x,y}$	0.0003	0.00001	30.000	$\ll 0.001$	0.99
	normalized pair frequency	$N_{x,y}/N_x$	0.1390	0.00160	86.875	$\ll 0.001$	0.76
	pair frequency	$N_{x,y}$	0.0390	0.00050	78.000	$\ll 0.001$	0.76
	loc. entropy	$H_s(x)$	1.8270	0.04300	42.488	$\ll 0.001$	0.65
	sched. entropy	$H_t(x)$	1.5860	0.08100	19.580	$\ll 0.001$	0.50
	sched. and loc. entropy	$H_{s,t}(x)$	2.5920	0.09600	27.000	$\ll 0.001$	0.61
cooperative	direct assists	$A_{x,y}$	0.1230	0.00100	123.000	$\ll 0.001$	0.98
	indirect assists	$V_{x,y}$	1.3170	0.01700	77.470	$\ll 0.001$	0.70
	betrayals	$B_{x,y}$	0.1460	0.00300	48.590	$\ll 0.001$	0.64

Table 3.1: Coefficients, $\hat{\theta}$, standard deviations, $\hat{\sigma}$, Z-scores, $|Z|$, p values, p, and AUC values for logistic regression models fitted to each individual feature for all friends and non-friends. AUC values of 0.5 correspond to a baseline random classifier.

average AUC scores of 0.830, 0.833, and 0.834 respectively. This similarity in performance is unsurprising considering the importance of temporal features (Fig. 3.2).

Surprisingly, fitting the model with just the cooperative features yields classification probabilities nearly as high (average AUC=0.789). This tree splits first on direct assistance $A_{x,y}$, in agreement with our expectation that latent friendship ties produce greater volumes of prosocial interactions than non-friend ties, followed by further splits on $A_{x,y}$ and indirect assistance $V_{x,y}$ over certain ranges of $A_{x,y}$. The fact that autocorrelation rather than direct assistance appears in the full model suggests first that autocorrelation is a more reliable indicator of latent friendship, but also that direct assistance may be capturing similar information. We test this idea by first training a classification tree using all features except autocorrelation $AC_{x,y}$. As expected, this tree splits first on high $A_{x,y}$, with the remaining structure being nearly identical to the models trained on all features or a subset, but substituting $A_{x,y}$ for $AC_{x,y}$. The average out-of-sample AUC for this set of trees is 0.800.

The structure and simplicity of the fitted trees suggest an underlying signature of friendship in the patterns of observed interactions. Specifically, highly periodic interactions

are strongly indicative of friendship because they require nontrivial levels of social coordination within the online environment. That is, friends must, and do, actively seek out each other in order to interact. Interestingly, although autocorrelation is highly predictive, combining it with spatial and schedule entropy reveals some subtleties in social interactions. When given all features or only temporal features, high autocorrelation $AC_{x,y}$ with high spatial and schedule entropy $H_{s,t}(x, y)$ yields a good predictor of latent friendships.[5] Entropy features by themselves are not particularly useful, but they do become predictive for high values of autocorrelation. Players with shared, low diversity playing habits (and thus low individual entropy levels) can appear in the data as synchronized, even without any social coordination. Entropy measures then allow us to identify non-friends who have autocorrelated schedules.

3.4.4 Lightweight predictors of friendship

These results suggest that individual features alone may perform well at predicting latent friendships, and such features would make good computationally lightweight predictors that could realistically be deployed on a large-scale system.

We explore this possibility using logistic regression to build single-feature latent tie classifiers and measure their performance using AUC. We divide our data into training and test sets using random partitions such that test and training sets are of equal size. [4] Figure 3.3 shows the ROC curves for each of these individual-feature models for predicting latent friendships, and the corresponding models are summarized in Table 3.1. Remarkably, the two most predictive individual features—autocorrelation $AC_{x,y}$ (temporal) and direct assistance $A_{x,y}$ (cooperative)—achieve near-perfect classifications, with AUCs of 0.99 and 0.98 respectively. To provide a comparison, we note that another method inferred friendship between graduate students with 96% accuracy using a single temporal-spatial feature [45].

[5] Note that while the classification tree only classifies friends and non-friends, the numbers observed, shown in the leaves of Figure 3.2, indicate the maximum likelihood estimates of friendship probability at the leaf.

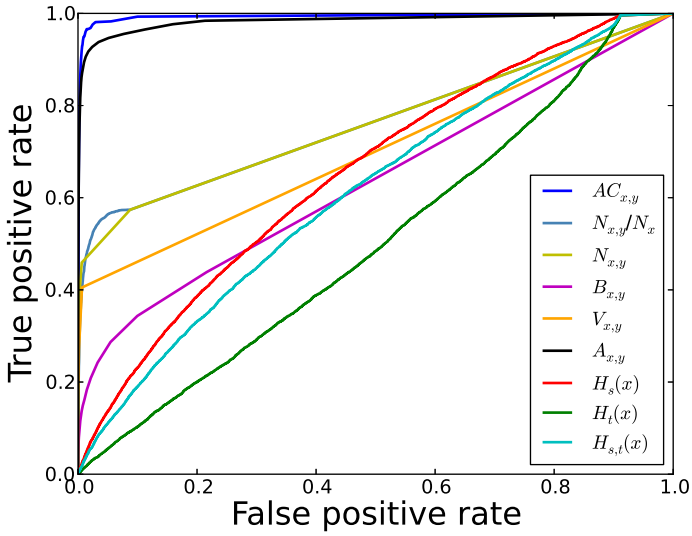

Figure 3.3: ROC curves for logistic regression models on individual temporal and cooperative features.

Both of our single-feature models are computationally lightweight and could thus potentially be deployed on a large-scale system to automatically infer latent ties for friendship-aware applications.

All of the remaining individual features perform more poorly, indicating that none would perform well as lightweight predictors in a real-world environment. Naïvely, we expected the volume of interaction $N_{x,y}$, and the fraction of that volume assigned to a particular other individual $N_{x,y}/N_x$, to be good indicators of latent ties. However, we find this not to be the case. Upon a closer examination of the mislabeled ties, we see that some latent ties spanned only a few interactions and this number was not significantly greater than the number of interactions with non-friends. Our autocorrelation feature is robust to this phenomenon because even these low-volume friendship ties exhibit strong periodicity in the interactions they generate.

Entropic features perform poorly alone because of insufficient diversity in location behavior within the population at large. That is, the number of interacting individuals at any

given time is large, while the number of "locations" is relatively small. As a result, both friend and non-friend pairs will often make similar choices about which locations to visit. Controlling for both time and space via $H_{s,t}(x)$ provides a narrower filter to individuals' behavior but does not substantially improve performance. Furthermore, our entropy measure does not consider the alignment of the individuals' schedules. As we saw with the classification trees, it is only in combination with other features, like autocorrelation, that entropy becomes predictive.

The failure of entropy features alone to perform well in *Reach* is interesting, and clarifies their success in applications to physical locations [36]. When the number of locations is large relative to the size of the population exploring them, the probability becomes very low that a non-friend pair will have similar distributions over locations in time. As the number of locations shrinks relative to the population size, this probability increases and eventually swamps the signal produced by friend pairs, which is what we observe in *Reach*. However, combining this signal with other features, like the autocorrelation, preserves some of its predictive power by mediating temporal effects with surprisingness, even in a system with densely occupied locations.

The poor performance of indirect assistance is unexpected, given that such behavior in *Reach* indicates a strong prosocial orientation and that direct assistance performs so well. Examining the mislabeled ties, we find that indirect assistance is not always possible in every interaction, i.e., in every game type, and even when it is possible, it is an uncommon event. These factors place tight constraints on its predictive power and the raw behavioral data we study contain examples of labeled friend pairs that exhibit no indirect assistance, thus making it difficult to identify a discriminative threshold.

Past work on friendship in *Reach* [80] suggested that our betrayal feature (in which an individual betrays their teammates to help their friends on the opposite team) should also correlate with latent friendship. And indeed it does: the average betrayal total $\langle B_{x,y} \rangle = 6.27$ for friend pairs but only 0.5 for non-friend pairs. The significance of this difference is qualified

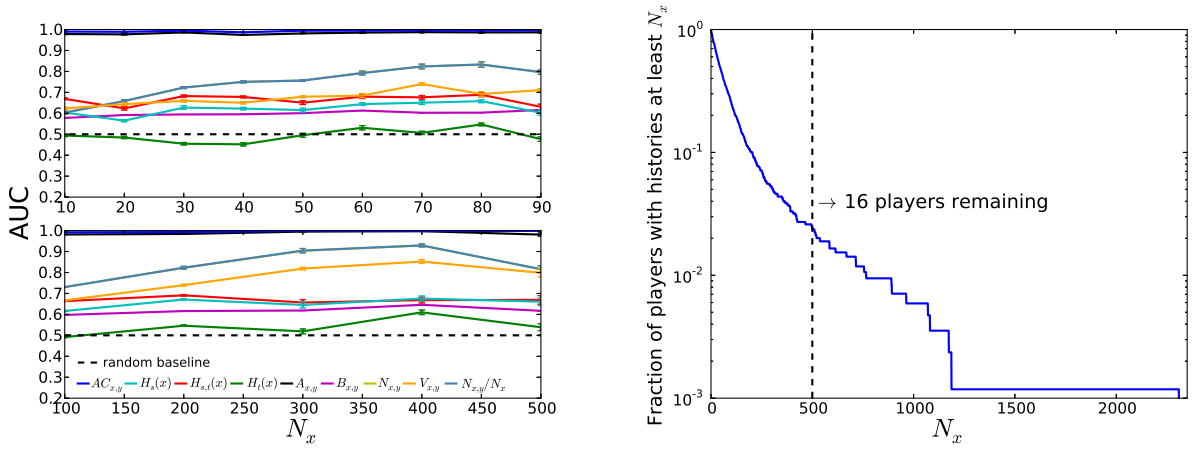

Figure 3.4: (Left) AUC as a function of N_x for each temporal and cooperative feature. The accuracy of $AC_{x,y}$ and $A_{x,y}$ are robust to available individual information while the accuracy of $V_{x,y}$, $N_{x,y}$ and $N_{x,y}/N_x$ increase with N_x. Entropic features remain relatively noisy regardless of N_x, see text for details. (Right) CCDF of N_x, number of games played, across all surveyed individuals.

by a substantially larger variance for friend pairs ($\sigma = 29.12$ versus 2.13), likely because many friends choose not to defect against their teammates, which lowers the discriminative power of this feature.

3.4.5 Predicting friendships for casual users

Achieving good predictions for the few users who produce large amounts of interaction data is useful. However, it is less useful if the performance degrades substantially as we consider users with progressively fewer observations, i.e., the casual users who typically make up the majority of individuals in an online system. To understand how robust our features are to the amount of available information, we study the performance of each individual feature as a function of N_x, the length of an individual's history.

We grouped surveyed individuals into bins according to the number of games they completed N_x. To provide a fine-grained look at individuals with short histories, where data are plentiful, and a coarse view of long histories, where data are sparse (Fig. 3.4, right), we

used bins of size 10 for $N_x < 100$ and bins of size 100 for $N_x \geq 100$.

We then computed the average AUC and its standard error by creating equal sized training and test sets from 10 random permutations of the data in each bin, and applying the individual-feature models. Examining these predictors' performance as a function of data volume provides some guidance for predicting friendships in data sets with large heterogeneities in data availability. Additionally, this test serves as a robustness check on our previous conclusions by implicitly considering the length of individual history as a feature.

Figure 3.4(Left) shows the average AUC for each feature as a function of history length N_x. Again two features, autocorrelation $AC_{x,y}$ and assists $A_{x,y}$, are consistently accurate predictors across all values of N_x. For the autocorrelation feature, this robustness indicates that pairs of friends interact more periodically than non-friends, regardless of their overall level of activity in the system. This signal is strong despite common individual schedules (e.g., weekend nights) that could potentially lead to artificially high autocorrelation between non-friends. Furthermore, even when an individual's data is sparse because he or she has completed very few games (less than 10), both autocorrelation and direct assistance have surprisingly strong predictive power, yielding average AUC values close to 0.98.

Focusing on autocorrelation, the reason for its high accuracy at small history lengths N_x is likely due to the large number of individuals in the system at any one time. This very large pool makes the probability very low for interacting with the same non-friend individual more than a few times. In real-world systems with low thresholds for two individuals meeting by chance (e.g., colocation in highly constrained or small physical environments), autocorrelation can be less discriminative and may require augmentation with other temporal or domain-specific features. Essentially, context can matter: it is unlikely that everyone who frequents the same busy coffee shop on Monday mornings will be friends, due to the nature of that location, while it would be a good bet that many pairs of individuals attending the same weekly soccer practice would be friends. The large effective capacity of an online system means that any signal from autocorrelation is likely to be significant.

In their analysis of friendship and gameplay in **Reach**, Mason and Clauset showed that individuals who are friends tend to coordinate and cooperate in ways that increase their team's score and the probability of winning the match [80]. The strongly predictive nature of direct assists $A_{x,y}$ that we observe corroborates this finding, and demonstrates that it holds over a wide range of N_x. That is, even for casual users, counting these prosocial interactions is a reliable indicator of friendship because friends do indeed cooperate more than non-friends.

Autocorrelation and direct assistance both maintain high performance across all sizes of N_x. The temporal features of raw and normalized pair frequencies $N_{x,y}$ and $N_{x,y}/N_x$ are less reliable predictors for small histories, but become more reliable as N_x increases. For large histories ($N_x > 400$), both features reach AUC values of nearly 0.90.

As we might have expected from our previous analysis, the performance of spatial and temporal entropy features $H_{s,t}(x)$, $H_t(x)$, and $H_s(x)$ do not improve as we accumulate more data. Similarly, we observe fairly weak improvements for indirect assists $A_{x,y}$ and betrayals $B_{x,y}$.

The remarkable accuracy achieved by our two best features, autocorrelation of schedules and direct assistance (prosocial interactions), demonstrate that lightweight predictors can be reliable even when applied to individuals with heterogeneous amounts of data by which to estimate latent friendships.

3.5 Social network inference

Given the excellent performance and computational efficiency[6] of the autocorrelation of co-play feature, $AC_{x,y}$, we use this lightweight predictor of friendship to infer the social network of the entire population of 17 million players. For each pair of players in the interaction network we compute $AC_{x,y}$, compare it to a threshold, which we explain below, and then label the pair of players as friends if their $AC_{x,y}$ is greater than or equal to the

[6] The autocorrelation function can be computed in $O(n \log n)$ time using a fast Fourier transform.

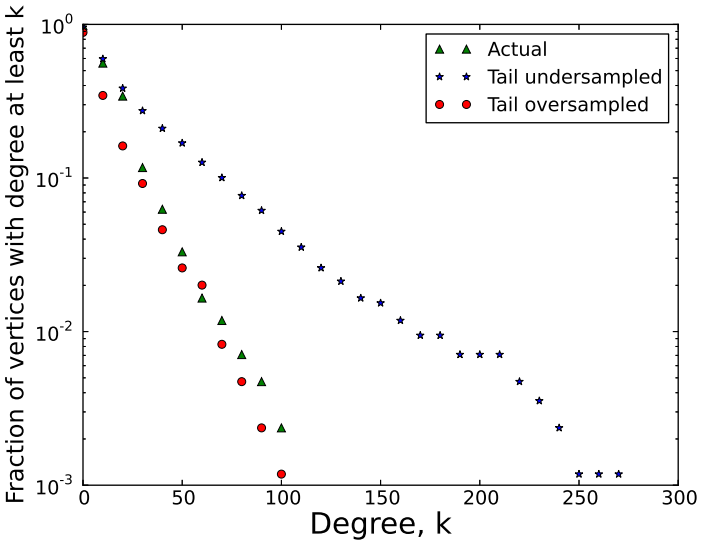

Figure 3.5: CCDF of actual and inferred degree distributions using only survey respondent data.

threshold value.

3.5.1 Threshold selection

The survey respondents are a biased sample of *Reach* players [80], being substantially more skilled than the typical player and investing roughly an order of magnitude more time playing than an average player. It is thus possible that the survey sampling bias has produced an oversampling or an undersampling of the tail of the degree distribution. In an attempt to control these opposing biases, we choose two thresholds, one to show what the network looks like if the survey respondents have less friends (undersampled tail) than the population, and one to show network structure if the respondents have more (oversampled tail).

Undersampled tail - To control for the undersampled tail bias we choose the $AC_{x,y}$ that

minimizes the Kullback-Leibler divergence

$$D_{KL}(P||Q) = \sum_i \ln\left(\frac{P(i)}{Q(i)}\right) P(i) \ , \tag{3.8}$$

where P is the degree distribution of social network derived from the survey respondent data and Q is the degree distribution calculated by creating edges between players x and y if their $AC_{x,y}$ is greater than or equal to a chosen threshold. As shown in Figure 3.5, this approach chooses $AC_{x,y} = 197$ and produces an inferred degree distribution for the entire network of 17 million players that matches the density near the head of the actual distribution but with a heaver tail than the survey data. It is not clear that this threshold choice necessarily produces an abundance of false friendships, as players with many friends are unlikely to have reported them all due to the tedious and time consuming nature of providing this information via the survey. This hypothesis is supported by empirical research, which showed that self-survey respondents tend to underestimate their interactions with individuals as a function of recency [45]. In our case, if a respondent did not interact with a friend recently, the tie may have been unreported.

Oversampled tail - To control for the oversampled tail bias, we compute the threshold by finding largest $AC_{x,y}$ that produces a degree distribution with a maximum degree no larger than the maximum degree observed in the survey. This approach chooses $AC_{x,y} = 1900$ and the tail of the inferred degree distribution agrees well with the survey data but less so near the head (see Figure 3.5).

3.5.2 Network structure

These two thresholds represent reasonable bounds for what we expect for our interaction data as a whole. We now apply these two thresholds to the interactions among the full 17 million players and study the structure of the induced social network. In the undersampled tail scenario ($AC_{x,y} = 197$), the inferred network consists of 8,373,201 nodes and 31,051,991

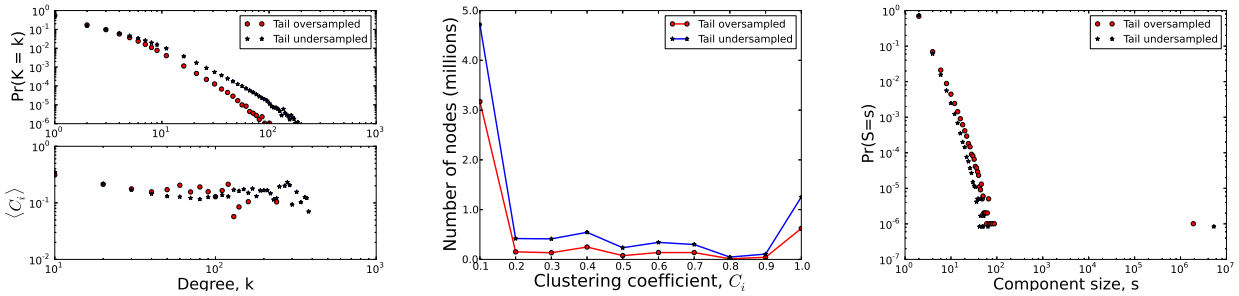

Figure 3.6: (Left) Degree distribution and mean clustering coefficient, $\langle C_i \rangle$ as a function of degree for both thresholds using the entire population of players. (Center) Binned clustering coefficient, C_i, plots for both thresholds using the entire population of players, bin width = 0.1. (Right) Distribution of component sizes. The undersampled tail network contains 1,194,032 components. The oversampled tail network contains 991,932 components.

edges, while the network inferred using the oversampled tail threshold ($AC_{x,y} = 1900$), contains 4,732,405 nodes and 11,435,351 edges.

The top panel of Figure 3.6(Left) indicates that both cases we observe degree distributions with heavy tails, where the majority of nodes in the network are connected to a small number of neighbors while a small number of nodes are connected to a large number of neighbors. When compared to the social graph of Facebook discussed in [115], players in *Reach* have smaller numbers of friends. The median friend count in Facebook is 99 while in *Reach* it is roughly $1/100^{th}$ the size, 1 and 2 at the over- and undersampled thresholds respectively. This large difference is likely caused by the high relative cost of establishing and maintaining a friendship in *Reach* versus the more cost-free nature of Facebook friendships. Specifically, *Reach* players must consistently and periodically interact over long periods of time, which is a significant investment of effort, while in Facebook, they must only click a request or accept button.

A vertex's clustering coefficient is defined as

$$C_i = \frac{\text{number of connected neighbors}}{\text{number of possible connected neighbors}} \ , \tag{3.9}$$

and provides a principled way of measuring how close vertex i and its neighbors are to

forming a clique [95]. This statistic equals unity when a vertex and its neighbors form a clique, while it equals zero when none of its neighbors are themselves pairwise connected. In our inferred graph, shown in the bottom panel of Figure 3.6(Left), a substantial fraction of individuals (between 16-20%) form tightly knit groups with high values of C_i.

Furthermore, the functional relation between the mean clustering coefficient $\langle C_i \rangle$ as a function of degree k_i is roughly the same, regardless of which threshold we choose (Fig. 3.6(Center)). For example, even when a vertex has a degree of 100, its clustering coefficient is likely to be between 0.1 and 0.2. This suggests that threshold choice does not substantially change the underlying network structure, and these numbers are close to those estimated for the Facebook social graph, where the mean clustering coefficient for a vertex with degree 100 was 0.14 [115]. While the mean clustering coefficient remains large independent of degree, a mild decreasing trend is evident. This suggests that nodes with high degree, who are likely high volume players, interact with others relatively less discriminately than nodes with smaller degrees, a pattern also found in the analysis of the Facebook social graph [115].

Figure 3.6(Right) plots the distribution of component sizes and indicates that the network contains a single large connected component composed of between two and four million players. The majority of the remaining nodes are spread amongst many components containing between roughly ten and twenty nodes. In the case of an undersampled tail, the network contains 1,194,032 components. In the oversampled case, the network contains 991,932 components.

3.6 Conclusion

Our motivating question was whether latent social ties like friendships can be accurately recovered from interaction data alone, and indeed we have shown that they can, with remarkable accuracy. We demonstrated that periodicity between interactions and specific prosocial behaviors across these interactions are both highly robust indicators of friendship, even in instances where data are sparse. Information theoretic measures of spatial and tem-

poral behavior, which are good indicators of the quantity of social ties in other contexts, are not effective at predicting the ties alone, but may be useful in combination with other temporal features. There are a number of interesting points these results suggest, both for improving *Reach* and for enabling friendship-aware applications in other domains.

Many online games, including *Halo: Reach*, rely on matchmaking algorithms to place individuals onto teams in order to make a new game instance go. If the *Reach* matchmaking algorithm works as desired, the teams are equally matched and the competition's outcome is unpredictable. However, when individuals play with friends, their performance improves [80], and this synergy is not included in the calculations of the matchmaking algorithm. A friendship-aware matchmaking algorithm, using features like the ones we consider here, could correct for the effective increase in team skill that occurs when friends play together, without reference to an external "friends list", and thus produce better matched teams, more enjoyable gameplay and overall greater engagement by the users. Another improvement would be to suggest as friends (to be added to a user's friends list) those individuals with whom a player has exhibited significant prosocial interactions, such as direct assists.

In the more general context of an online system where we can observe interactions, but not labeled friendship ties, our results could be applied in an unsupervised manner. Using an unsupervised learning algorithm such as k-means to separate friends from non-friends based on the autocorrelation values of their co-interaction time series should be relatively simple and robust. The discriminatory power of autocorrelation and prosocial behavior, even with sparse data, suggests that latent friendship ties may in fact be easily detectable, due to the nature of friendship itself. In a sense, periodic and prosocial interactions are the definition of friendship, and it may be difficult to maintain such a relationship online without manifesting a signal in these ways.

Friendship-aware applications are only one new opportunity presented by the automatic inference of latent social ties from interaction data. The ease with which we were able to

recover the latent friendship labels raises significant privacy questions, as these labels are often considered private information. The accurate recovery of such private signals from public interaction data may facilitate malicious applications. The social consequences of large-scale deployment of friendship inference is difficult to estimate.

Other benefits are more easily identified. For instance, many questions in computational social science may benefit from the accurate recovery of the underlying social network that generates the observed data. The general outlines of our results may have productive applications in many of these domains, e.g., in big data analyses of online social behavior. Our results are encouraging for settings where ground-truth data are at best rare and expensive to collect. Robust methods to extrapolate from ground-truth survey data to large-scale latent social network prediction are of great practical interest. We look forward to seeing the exploration of these and other beneficial applications.

Chapter 4

Social Network Dynamics in a Massive Online Game: Network Turnover, Non-densification, and Team Engagement in Halo Reach

4.1 Introduction

Although social networks are inherently dynamic, relatively few studies have analyzed the dynamical patterns observed in large, real-world, online social networks [1] . Insights into the basic organizing principles and patterns of these networks should inform the development of probabilistic models of their evolution, the development of novel methods for detecting anomalous dynamics, as well as the design of novel, or augmentation of existing, online social systems so as to support better user engagement. Here, we present a brief empirical analysis of a novel online social network, and use its dynamics as a lens by which to both identify interesting empirical patterns and test existing claims about social network dynamics.

This network is drawn from a massive online game *Halo: Reach*, which includes the activities of 17 million unique individuals across roughly 2,700,000 person-years of continuous online interaction time. Multiplayer games like these are a highly popular form of online interaction, and yet, likely due to the lack of data availability, have rarely been studied in the context of online social networks. As a result, we know more about networks like Facebook and Twitter, which consume relatively little time per user per week, than we do about more immersive online social systems like multiplayer games, which consume on average more than

[1] This chapter was published as S. Merritt, A. Clauset, Social Network Dynamics in a Massive Online Game: Network Turnover, Non-densification, and Team Engagement in Halo Reach, Workshop on Mining and Learning from Graphs, Aug. 2013

20 hours of time per user per week [6].

Furthermore, online games represent a large economic sector, including not just the sale of the software itself, but an entire ecosystem of entertainment products, worth billions of US dollars worldwide. For example, on the first day of sale *Halo: Reach* grossed nearly $200 million in revenue from roughly 4 million copies of the game. Increasingly, one point of attraction to this and other online games is their multiplayer component: while it is possible to play individually, the online system is designed to encourage and reward social play. Despite their enormous popularity, relatively little is known about the social networks within these online games, how they evolve over time, and whether the patterns they exhibit agree or disagree with what we know about online social interactions from other online systems.

This raises two interesting questions. First, what is the shape of a social network in a massive, multiplayer online game? And second, how does this network change over time? Answers to these questions will shed new light on the relative importance of a game's social dimension and on the relationship between a game's social structure and its long term success, and will clarify the generality of our current beliefs about online social networks, which are largely derived from non-game systems like Facebook, Twitter, Flickr, etc.

One immediate difference between the social network underlying online games and those more classically studied is the competitive nature of games. The large scale of and rich data produced by online games provides a novel perspective on certain other social interactions, e.g., team competition. Past work has shed light on competitive dynamics [81], social organization [112], economic trading networks [69], and deviant behavior [19]. Here, we study the underlying friendship network from *Halo: Reach*, which we infer [84] from 18 billion interactions between 17 million individuals over the course of 44 weeks. Edges in this network represent inferred online or offline friendships, in contrast to mere online interactions alone, which the Halo system generates randomly via the matchmaking system that places players into competitions.

We first analyze the static (cumulative) friendship network and find that it contains a heavy-tailed degree distribution, which appears more log-normal than power law. This network is composed of many small disconnected components and a single giant component containing roughly 30% of all players. A plurality of vertices exhibit low clustering coefficients ($0 < c_i \leq 0.1$), indicating very sparse local structure, while the next-most-common pattern is to exhibit a very high clustering coefficient ($0.9 < c_i < 1.0$), indicating very dense local structure.

To study the friendship network's evolution we create a time series of 44 network snapshots, one for each week of the year. Initially, this network exhibit strong structural turnover and decays rapidly from a peak size to a more stable, but dynamic core. Despite its apparently stable size, we find steady network turnover over most of the time period, indicating a dynamic equilibrium as roughly equal numbers of vertices join and leave the network each week. Furthermore, in contrast to other online social networks [72], the Halo friendship network does not densify over time: the mean degree and the average pairwise distance within the giant component are stable.

Finally, we observe that the network's clustering grows over the first 25 weeks and then declines, suggesting that players form increasingly tight knit groups during the first half of the network's life only to disband them later. In addition, we find a positive relationship between the consecutive weeks played and the player's mean clustering coefficient. This indicates that players who belong to tightly knit groups tend to also play longer and more consistently.

The remainder of this paper is organized as follows: first we discuss related work. Next, for completeness, we introduce the data and the results of an anonymous online survey. Following this, we briefly describe the method used to infer the social network. These topics are described in detail in [84]. Following this, we analyze the inferred social network statically and then dynamically. We conclude with a discussion and final thoughts on future work.

4.2 Related work

Work related to the analysis of networks began with the study of static networks. Researchers identified important structural patterns such as heavy tailed degree distributions [13], small-world phenomenon [86, 119], and communities, as well as probabilistic models and algorithms that produce and detect them [94, 32].

As new online social systems continue to emerge on the web, the static analysis of social networks continues to be an area of great interest. Researchers have a steady stream of new empirical network data with which they can test new and existing theories about social dynamics, whose sources include Twitter [73], Facebook [115], Orkut and Flicker [87].

More recently, researchers have begun to study how time influences [30] and changes network structure [3]. New dynamical patterns, such as densification [75, 72] and shrinking k-cores [52], as well as probabilistic models have been identified that shed light on how changes in the underlying processes that produce these networks affect its structure.

Since the underlying processes that produce complex networks are typically not random, a rich body of work related to mathematically modeling various networks whose structure cannot be generated by an Erdős-Rényi random graph model has emerged [48]. One such model is the configuration model, which produces a graph with a predefined degree distribution by randomly assigning edges between vertices according to their degree sequence [96]. Another is the preferential attachment model, which is a generative model that produces heavy tailed degree distributions by assigning edges to vertices according to the notion of "the rich get richer". That is, edges are assigned to vertices according to how many they already have. Other approaches that produce specific properties, such as short diameters and communities include the Watts and Strogatz model and stochastic block model respectively [119, 54].

Lastly, we would be remiss if we did not also mention work related to the study of online games. Most uses of online game data have focused on understanding certain aspects

of human social behavior in online environments. Examples include individual and team performance [107, 108, 105, 106], expert behavior [65, 62], homophily [63], group formation [64], economic activity [29, 10], and deviant behavior [2]. Most of this work has focused on massively multiplayer online role playing games (MMORPGs), e.g., World of Warcraft, although a few have examined social behavior in first person shooter (FPS) games like *Reach* [105]. Relatively little of this work has focused on the structure and dynamics of social networks.

A particularly unique aspect of our work is our data set. Few studies have analyzed the temporal dynamics of a social network derived from a popular form of social media, online games. Moreover, the majority of studies that examine the temporal dynamics of networks only analyze networks that primarily grow over time. That is, once a vertex is added to these networks it is never removed. In our network, vertices enter and leave the network consistently over time.

4.3 Data and survey

4.3.1 Game details

Our data are derived from detailed records of game play from *Halo: Reach*, a popular online first person shooter game. Individual game files were made available through the Halo Reach Stats API.[2] Through this interface, we collected the first 700 million game instances (roughly 305 days of activity by 17 million individuals). Among other information, each game file includes a Unix timestamp, game type label, and a list of gamertags. This large database provides us with complete data on the timing and character of interactions between individuals but provides no information about which interactions are produced by friendships versus non-friendships.

[2] The API was active from September 2010 through November 2012. API documentation was taken offline in September 2012.

4.3.2 Survey

We combine these in-game behavioral data with the results of an anonymous online survey of *Reach* players [80]. In the survey, participants supplied their gamertag from which we generated a list of all other gamertags that had ever appeared in a game with the participant. From this list, the participant identified which individuals were friends. [3] We interpret these subjective friendship labels as ground truth. From these data, we constructed a social network with links pointing from participants to their labeled friends. In our supervised learning analysis, both a labeled friendship and the absence of a label are treated as values to be predicted (i.e., we assume survey respondents explicitly chose not to label their co-player as a friend). Of the 965 participants who had completed the friendship portion of the survey by April 2012, 847 individuals appear in our data (the first 305 days of play); this yielded 14,045 latent friendship ties and 7,159,989 non-friendship ties.

4.3.3 Interaction network

We represent the set of pairwise interactions as a temporal network, in which edges have endpoints and exist at a specific moment in time. Vertices in the network correspond to gamertags, and two vertices are connected if they appear in a game instance together at time t (time of day, in 10 minute intervals). Each vertex thus has a sequence or time series of interactions with other vertices. The resulting network, derived from our complete game sample, contains 17,286,270 vertices, 18,305,874,864 temporal edges, and spans 305 days. The subgraph of interactions by our survey participants contained a total of 2,531,479 vertices and 665,401,283 temporal edges over the same period of time.

[3] In the survey, a friend is defined as a person known by the respondent at least casually, either offline or online.

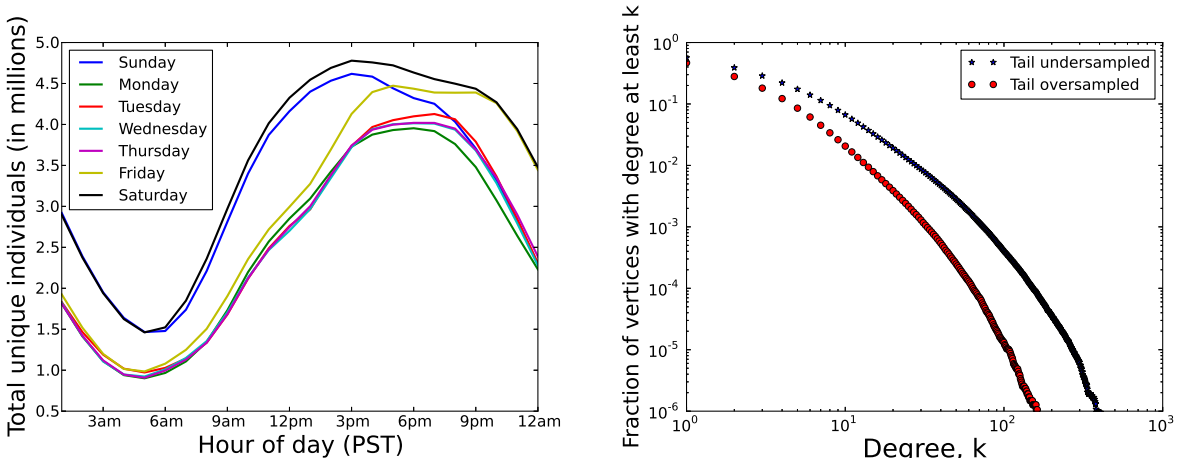

Figure 4.1: (Left) Number of unique individuals ever seen at a given time of day (in Pacific Standard Time), across the 305 days spanned by the data, illustrating significant daily and weekly periodicities. (Right) Complementary cumulative distribution functions of vertex degrees

4.4 Inferring the social network

Pairs of individuals in *Reach* that are friends are known to play many more consecutive games (12, on average, or about 2 hours of time) than non-friends (1.25, on average) [80]. Thus, continuous interaction over a significant span of time is likely an indication of a latent tie, while more intermittent interactions likely indicate a non-friend tie, given the large population of non-friends available to play at any time. The expected diurnal and weekly cycles observed in the data will modulate these behaviors, and a reasonable approach for their quantification is via interaction periodicity (see Fig. 4.1, Left). Let

$$n_{x,y}(t) = \mathbb{1}\{x \text{ and } y \text{ play together at time } t\} \tag{4.1}$$

represent the time series of binary interactions between individuals x and y, where 1 indicates an interaction at time t and 0 indicates no interaction. If x and y are friends, we expect $n_{x,y}(t)$ to exhibit stronger periodicity than for non-friends. This expectation may be quantified as

the autocorrelation of the time series $n_{x,y}(t)$ over all time lags τ:

$$AC_{x,y} = \sum_{\tau} \sum_{t} n_{x,y}(t) n_{x,y}(t-\tau). \tag{4.2}$$

If $n_{x,y}(t)$ is generated by a non-friend pair, $AC_{x,y}$ should be small because these individuals do not interact regularly. On the other hand, if $n_{x,y}(t)$ is generated by a friend pair, we expect $AC_{x,y}$ to be large. Training a logistic regression classifier on this feature alone correctly identifies 95% of ties, even for casual users (i.e. those that play few games). For a detailed presentation of this analysis, see [84].

The survey respondents are a biased sample of *Reach* players [80], being substantially more skilled than the typical player and investing roughly an order of magnitude more time playing than an average player. It is thus possible that the survey sampling bias has produced an oversampling or an undersampling of the tail of the degree distribution.

In an attempt to control these opposing biases, we choose two thresholds, one to show what the network looks like if the survey respondents have less friends (undersampled tail) than the population, where $AC_{x,y} = 197$, and one to show network structure if the respondents have more (oversampled tail), $AC_{x,y} = 1900$. Details of the methods used to select these thresholds can be found in [84].

4.5 Network structure

The two thresholds computed using the survey data in the previous section represent reasonable bounds on what we expect to observe in the data at large. In this section, we use both thresholds to analyze the structure of the inferred social network and show that network structure remains invariant to threshold choice. In the undersampled tail scenario ($AC_{x,y} = 197$), the inferred network consists of 8,373,201 nodes and 31,051,991 edges, while the network inferred using the oversampled tail threshold ($AC_{x,y} = 1900$), contains 4,732,405 nodes and 11,435,351 edges.

Figure 4.1(right) plots the complementary cumulative distribution of vertex degree

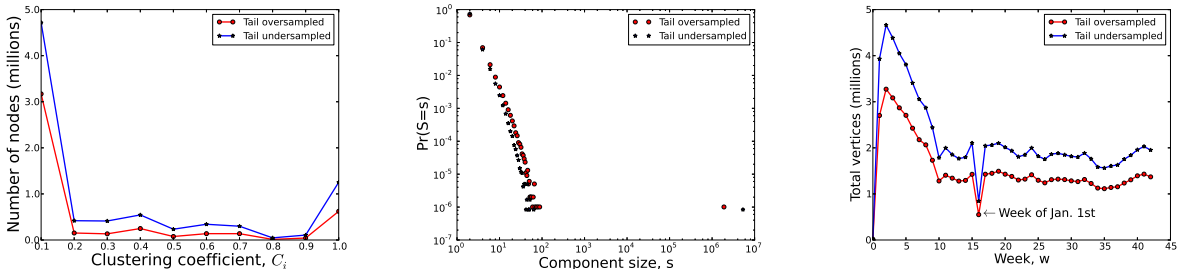

Figure 4.2: (Left) Histogram of vertex level clustering coefficients, C_i (Center) Distribution of component sizes. (Right) Total number of vertices in the network as a function of week, w. All figures plot distributions for networks using under and over sampled tail thresholds ($AC_{x,y} = 197$ and $AC_{x,y} = 1900$ respectively)

sizes and indicates that the network is primarily comprised of vertices with only a few edges and only 10% containing more than roughly 10 or 20, depending on threshold choice.

To quantitatively measure the degree to which groups make up the network, we compute the vertex level clustering coefficient, defined as,

$$C_i = \frac{\text{number of connected neighbors}}{\text{number of possible connected neighbors}}. \tag{4.3}$$

C_i provides a principled measure of how close vertex i and its neighbors are to forming a clique [95]. A clustering coefficient equal to one indicates the vertex and its neighbors form a clique, while a coefficient equal to zero indicates none of the vertex's neighbors are connected.

In addition to vertices containing only a few edges, the majority possess low clustering coefficients, indicating that most players usually choose to play games with only one other person at a time (see Fig. 4.2, left). There is also a small, but non-trivial group of players who have high clustering coefficients, indicating that they choose to play with two or more friends, who are also friends themselves.

Figure 4.2(center) plots the distribution of component sizes and indicates that the network contains a single large connected component composed of between two and four million players. The majority of the remaining nodes are spread amongst many components containing between roughly ten and twenty nodes. In the undersampled tail case ($AC_{x,y} = 197$), the

network contains 1,194,032 components. While in the oversampled tail case ($AC_{x,y} = 1900$), the network contains 991,932 components.

From this analysis, we can conclude that the network, as a whole, contains between 23%-47% of the total 17 million player population and that these players tend to interact with their friends in pairs or small densely connected groups.

4.6 Network dynamics

The static analysis in the previous section sheds light on the overall structure of the social network and provides clues about how friends interact in the game, but it provides no insight into how the network changes over the course of its 44 week time span.

To study the friendship network's evolution, we create snapshots of the network by extracting the edges of inferred friends from the interaction network at weekly time intervals, where each interval is aligned with a week of the year. The set of snapshots we study begins on the 37^{th} week of September, 2010, and extends through the week 28^{th} week of 2011, totaling 44 weeks of time.

4.6.1 Players

In this section, we study how the network's population evolves over time. Figure 4.2(right) plots the inferred number of players in the social network, under both thresholds, over each week in the data. On launch week, we observe a small population that quickly grows to a peak of roughly 3.25 million and then decreases linearly to roughly 1.3 million. This pattern suggests that gamers initially play the game alone, then transition to a social mode of play[4] , and later reduce their volume of overall play. The large decrease in population can likely be attributed to players becoming bored with the game. That is, once the initial excitement and novelty of the game wears off, players spend less time playing the game.

[4] **Reach** sold nearly 4 million copies of the game on the first day of sale. (see `http://en.wikipedia.org/wiki/Halo:_Reach`)

Following the population decrease in weeks 3-10, the number of vertices remains relatively constant until week 16, the week of January 1^{st}, where a rapid, but temporary drop, in the total population occurs. One possible cause of this drop in population could be due to an Xbox Live service outage, although we find no record of such an event on the web.

Over the course of the remaining weeks, the population slowly and mildly declines, reaching a minimum during the month of June, and then returning to the levels seen in week 10. This weak decline and growth pattern is likely due to players having to spend increasing amounts of time away from the game studying for end of semester/year exams, since the majority of players are between the ages of 16 and 24, as noted in [80]. This also explains the growth in population following the month of June, when nearly all schools have ended for the year.

4.6.2 Network Turnover

Next, we study the rates at which players enter and exit the social network over time. Figure 4.3(top-left) plots the cumulative fraction of observed players as a function of week, w. Based on the dynamics observed in Fig. 4.2(right), it is not surprising to see that within the first 10 weeks of play, roughly 60% of the network's population has appeared in the network at least once. One might expect the population to grow exponentially, as observed in weeks 0-10, and then become constant, indicating that a large and constant number of players enter the network early and then consistently re-appear week after week. However, the dynamics that play out following week 10 indicate that the remaining 40% of the population continues to enter the network at a nearly constant rate. This linear growth may indeed be due to new players entering the social network, however, it may also be due, in part, to existing players changing their gamer tags, a relatively simple and inexpensive transaction (a player may change their gamertag once for free).

The relatively constant size of the network after week 10 and the constant inflow of new players suggests that an equally sized portion of the population also exits the network

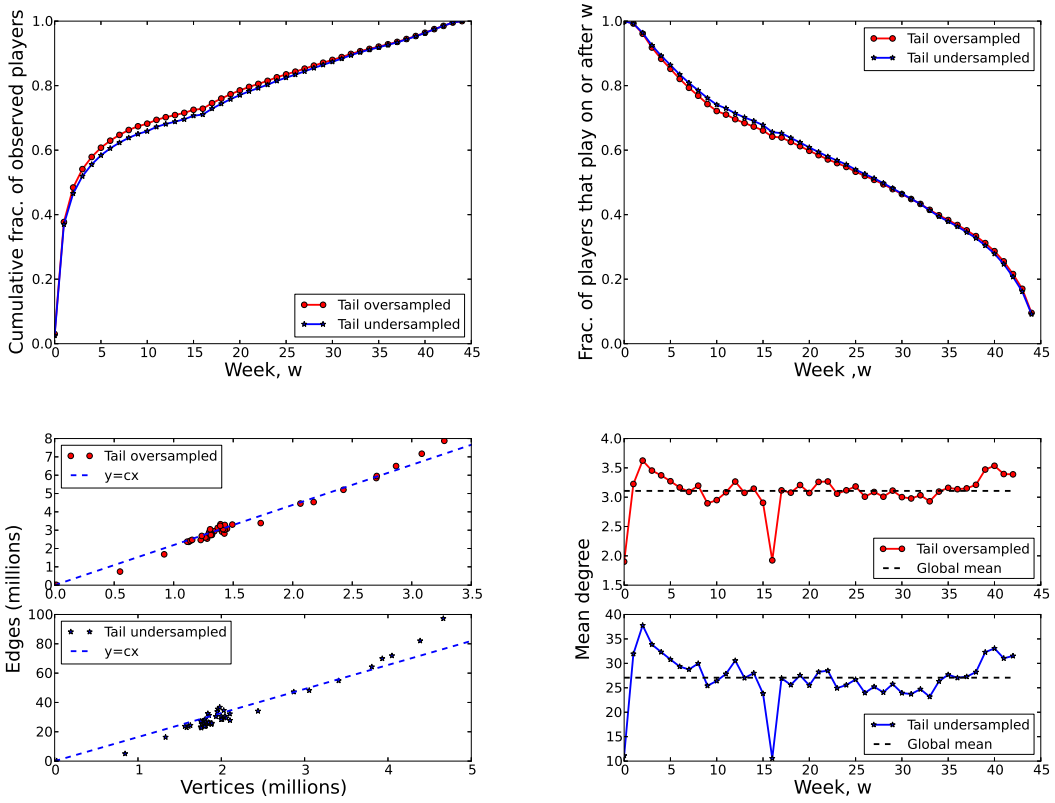

Figure 4.3: (Top-Left) Cumulative fraction of observed players per week, w. (Top-Right) Fraction of players in week w that play on or after week w. (Bottom-Left) Number of edges as a function vertices and linear fit (oversampled tail $r^2 = 0.99$, $c = 2.188$, $p \ll 0.001$, undersampled tail, $r^2 = 0.98$, $c = 16.389$, $p \ll 0.001$), indicating that the network is *not* densifying. (Bottom-Right) Mean degree as a function of week, w

over time, producing a "dynamic equilibrium". Indeed, Figure 4.3(top-right) plots the fraction of players that reappear at least once on or after week, w. While the dynamics in Figure 4.2(right) indicate players may enter the network and play a large quantity of games with friends and then exit, we find this not to be the case. The steady, nearly linear decline in the number of players that re-appear in future network snap shots after week 10 indicates that players are not quick to leave the game permanently; a pattern that is likely the result of players establishing consistent schedules with their friends.

4.6.3 Network non-densification

Next, we examine how the density of the network varies with its size. In many networks, including social networks, it has been shown that a network densifies with the number of vertices [75, 72]. That is, the number of edges in a network grows super linearly with the number of nodes. However, the friendship network studied here does not exhibit this pattern. As shown in Figure 4.3(bottom-left), the number of edges in the network grows linearly with the number of the vertices under both thresholds. A super linear relationship between vertices and edges would suggest that as the network grows in size, players would acquire increasing numbers of friends.

The non-densification of the network is also expressed in Figure 4.3(bottom-right), which plots the network's mean degree as a function of week, w. The mean degree of vertices remains relatively constant over time, with the exception of week 16, further demonstrating that the network does not densify over time.

The linear relationship between vertices and edges in the network should come as no surprise. Establishing a link in **Reach** comes at a high social cost. Players must spend considerable time coordinating when and how often to play. Additionally, players must invest significant portions of time interacting with one another over the course of many games. These conditions make it difficult and time prohibitive for players to establish increasing numbers of friendships as the network grows, even though many may be available. This pattern contrasts with other online social networks, where creating an edge is as simple and cheap as accepting a friend request [72].

This result indicates network densification is not a universal property of dynamic networks and that the cost of establishing links is likely an important underlying mechanism which controls the property. Thus, we should expect to find densification patterns in networks where linking is cheap and non-densification where establishing a link is expensive.

4.6.4 Giant component

Like many other social networks, the **Reach** friendship network is composed of many disconnected components of varying sizes (see Fig. 4.2, right). Here, we study the dynamics and connectedness of the network's giant component, which contains between roughly 80-90% of the network's vertices and compare it to random networks generated via the Configuration Model, using the degree sequences of each network snapshot [96].

The geodesic distance between vertices is defined as the shortest path connecting them [95]. We estimate the mean geodesic distance, also known as the average diameter, of the network at each week, w, by taking the mean value of the geodesic distances of 1,000 randomly selected pairs of vertices from the giant component.

As shown in Figure 4.4(top-left), we find the distance between vertices in the giant component to be relatively constant, with the exception of week 16, and equal to roughly 10 or 4.5 for over and under sampled tail thresholds respectively. This constant, **non-shrinking** diameter pattern contrasts with observations in citation [75] and other social networks [72].

The mean pair-wise distances of the friendship networks are also larger than that of the Configuration Model based random networks. This is not surprising since random graphs have no clustering. That is, in the Configuration Model networks, edges are placed between nodes at random, thereby creating a giant component with a smaller diameter, whereas in the friendship networks, edges tend to be placed within groups rather than between them, inducing a giant component with a larger diameter. Additionally, and as expected, the friendship network's clustering structure creates giant components of smaller size, when compared to the Configuration Model graphs (see Fig. 4.4, top-right). That is, the clustering structure produces more disconnected groups of vertices in the friendship network than in the Configuration Model.

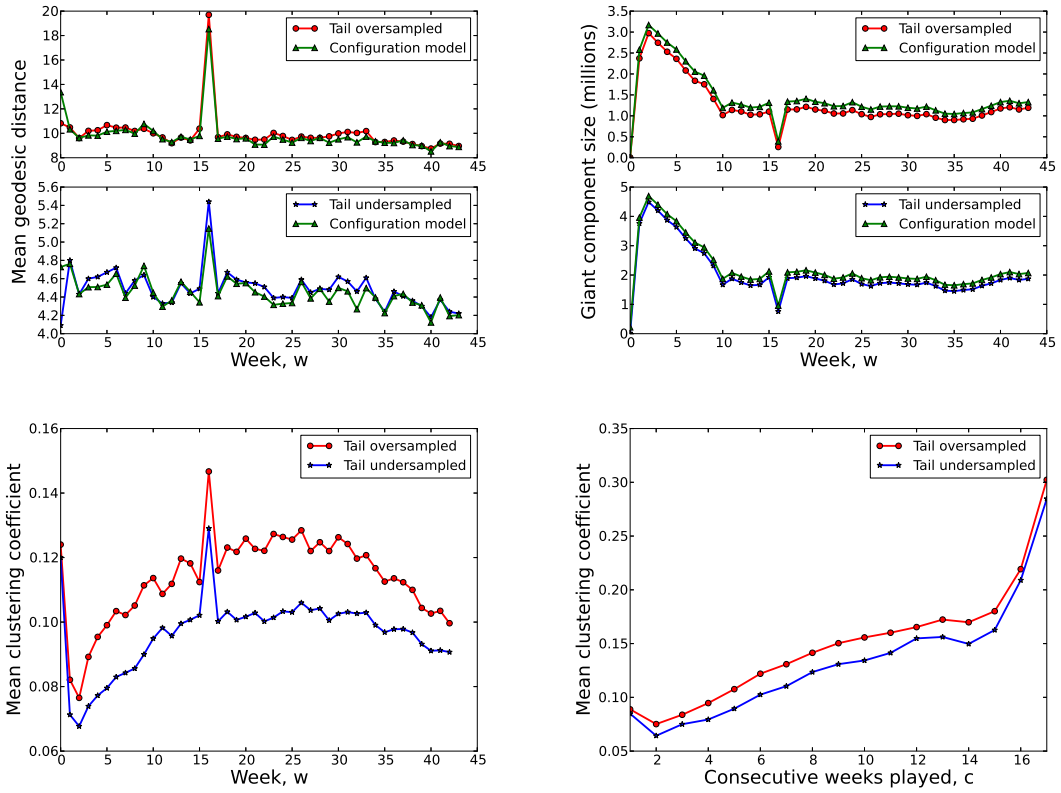

Figure 4.4: (Top-Left) Mean geodesic distance between 1000 randomly selected pairs of vertices in the network's giant component (Top-Right) Giant component size as a function of week, w. (Bottom-Left) Mean clustering coefficient as a function of week, w. (Bottom-Right) Mean clustering coefficient as a function of consecutive weeks played, c

4.6.5 Group dynamics

Recall, that a non-trivial fraction (between 16-20%) of the overall population in *Reach* is composed of tightly knit groups (see Fig. 4.2, left). Here we study how clustering amongst players evolves over time.

Figure 4.4(bottom-left) plots the mean vertex level clustering coefficient over the course of the data's 44 week time span. During the first week, the network is clustered nearly as much as during its peak. This indicates the first week's small population of players are more social. That is, they play more in groups than the players in the immediately following weeks. After week 1, we observe a parabolic trajectory of the mean vertex level clustering

coefficient. This pattern indicates that, through week 25, an increasing fraction of players tend to interact in increasingly connected groups. After this peak, the clustering within the network decreases, indicating these groups deteriorate and players choose to interact more in a pairwise fashion.

We also study how clustering correlates with play habits. For each player in the network, we calculate the largest consecutive number of weeks the player appears in the data. For each set of players playing c consecutive weeks, we compute the mean clustering coefficient, calculated from the static graph, i.e., the union of all 44 weekly graphs.

Figure 4.4(bottom-right) plots the mean clustering coefficient as a function of consecutive weeks played and indicates a positive correlation between the number of consecutive weeks played and clustering coefficient (oversampled tail $r^2 = 0.85$, $p \ll 0.001$, undersampled tail $r^2 = 0.81$, $p \ll 0.001$). This pattern indicates that socially engaged players, who are members of increasingly connected groups, are retained and engaged in the system for longer and more consistent periods of time.

4.7 Conclusions

Here, we studied the evolution of a novel online social network within the popular online game *Halo: Reach*, which we inferred from billions of interactions between tens of millions of individuals. We find an interesting two-phase pattern in the structural turnover of the graph, with a large "churn" in the beginning, as individuals try out the system briefly and then leave, followed by a more prolonged "dynamic equilibrium" period, characterized by a stable-sized giant component with roughly equal rates of vertices joining and leaving. Furthermore, we find that the friendship network does not densify over time and its diameter does not shrink, in contrast to other online social networks. This particular pattern is likely attributable to the high-cost of friendship links in this network, which require genuine and prolonged investment of time in order to maintain. This "high cost" requirement contrasts with the low- or zero-cost of maintaining links in most other online social networks. As a

result, it seems likely that other online social networks with high costs for link formation and maintenance would also exhibit non-densifying patterns.

One relatively understudied aspect of online social network structure is the behavior of small groups of friends. In Halo, these groups have a functional purpose, as they constitute a coherent "team" of players that engage the system together. We find that the local network density (cluster coefficient) amongst individuals tends to increase over the first 25 weeks of our study period and then declines. Additionally, we find that a player's clustering coefficient is positively correlated with consecutive numbers of weeks played. This indicates players who are more socially engaged within the game are also more engaged in game play itself.

The long term financial success of many online games, including **Reach**, rely not only on selling millions of copies of the game, but also on retaining players through the sale of subscription based online services. We observe that players who remain in the social network the longest are members of tighter knit groups than others. This suggests that the social aspect of **Reach** has a strong influence on the play patterns of its members. That is, the level of local social engagement with friends appears to correlate strongly with long-term engagement with the game itself.

One interesting question for future work is whether the overall efficacy of the game's matchmaking algorithm can be improved by explicitly accounting for the synergistic effect of playing with friends. That is, providing additional opportunities for friendship to engage socially with each other may facilitate the overall engagement of more weakly connected social groups, who may otherwise disengage the game earlier than desired.

In short, the parabolic structure of the network's clustering suggests that groups may form and disband over time. Studying this pattern in more detail and understanding its effects on other play patterns, such as its influence on competition outcomes and participation in different types of competitions, would shed light on how group dynamics applies to online social systems and how group skill and preferences evolve over time.

Finally, the temporal patterns observed in the Halo friendship graph demonstrate the

utility of online games for studying social networks, shed new light on empirical temporal graph patterns, and clarify the claims of universality of network densification. We look forward to future studies exploring other dynamical aspects of the Halo data and the social networks embedded in other online games.

Chapter 5

Environmental structure and competitive scoring advantages in team competitions

5.1 Introduction

Professional team sports are a rich and relatively controlled domain through which to investigate fundamental questions in both the dynamics within and across competitions between groups, and the factors that determine competitive outcomes [99, 67] [1] . With many possible actions and many possible payoffs, such games are a kind of dynamical competition [51], in contrast to the strategic interactions of classic game theory [88]. A distinguishing feature of most such competitions is their structurally homogeneous or "level" playing field, which allows differences in team scores to be attributed to one team being relatively more skilled than another, or, if the difference is small, to chance events [43, 60].

It thus remains unknown what impact structural heterogeneities, like an irregular playing field, variations in rules, or differences in resources, may have on a competition's internal dynamics. Heterogeneities may produce structural competitive advantages [14], allowing a team to perform above its skill level by exploiting these environmental irregularities. In fact, the roles in shaping competition dynamics and outcomes of skill, structure, and chance remain highly controversial, both in sports [15] and in other types of social competition [14, 103, 44]. A better understanding of these principles would inform the design of novel competitive environments [85, 97], and could shed light on competition dynamics

[1] This chapter was published as S. Merritt, A. Clauset, Environmental structure and competitive scoring advantages in team competitions, Scientific Reports, Oct. 2013

in other domains, such as ecology and evolutionary biology [21], political conflict [88] and economics [22].

Online games present a novel approach to investigate these questions. Such games encompass a broad and growing variety of relatively controlled competitions, played by hundreds of millions of individuals [47] and producing large quantities of detailed observational data. We study a unique data set drawn from the popular online game *Halo* (see *Appendix A*), a kind of virtual team combat, which contains nearly 1 billion scoring events across roughly 10 million diversely structured team competitions. Each of these competitions is roughly independent, such that team memberships are substantially randomized and no acquired resources are carried to the next competition. This property thus mitigates the confounding effects of cross-competition correlations present in professional sports and allows us to study how structural variations shape competition dynamics and outcomes.

We partition these competitions according to their particular environmental structure, competition rules, resource quality and difference in team skill, and characterize their scoring dynamics via a probabilistic model. The resulting model parameters provide a compact representation of the associated competitive dynamics, and serve as targets to be explained by variation in a competition's structural features.

Despite wide variation, structure has a modest impact on the tempo of events, but a large impact on the scoring balance, i.e., the difference in team scores. Additionally, the rate of scoring events over time exhibits the same three-phase pattern observed in professional sports [50]. Overall, structural features alone are highly predictive of overall competition tempo, the range of competitive scoring advantages, and ultimate predictability of the competition's outcome. Like business firms competing in the marketplace [14], teams generally exploit environmental and resource heterogeneities for sustained competitive advantage. However, contrary to the pattern of professional sports, the most balanced competitions—those with narrow margins of victory—arise from specific environmental heterogeneities, not from equally skilled teams competing in homogeneous environments. These results illustrate

the rich potential of online game data for investigating social dynamics and competition [112], clarify the role of chance when teams are well matched, and point to specific design principles for balanced competitions.

5.2 Modeling Competition Dynamics

We first introduce the notion of an "ideal" competition, in which perfectly matched teams play on a level field with no exploitable features. Such a competition's outcome is thus determined solely by the occurrence and accumulation of chance events, e.g., accidents, miscalculations, and events outside direct control. In this way, the highly strategic and carefully motivated actions of equally skilled teams will effectively produce purely stochastic dynamics.

These dynamics can be described by a particularly simple stochastic process [27]. Scoring events occur infrequently and independently, and their pattern follows a Poisson process with rate λ_0—a common assumption in quantitative analysis and modeling of professional sports [50, 113, 61]. Given a scoring event occurs, a fair coin determines which team accrues points from it. The difference in scores between teams thus follows an unbiased random walk, and scoring overall follows an equiprobable or balanced Bernoulli scheme.

Real competitions, with heterogeneous structure or skill differences, will deviate from this ideal. We capture these deviations through a generalized model, which may be fitted directly to scoring data and whose parameters quantify the size and character of the non-ideal patterns. We then investigate the extent to which the observed non-ideal patterns can be predicted from variation in competition structure.

We assume a competition between teams r and b, and we let $s_r(t)$ denote team r's cumulative score at an intermediate time $t < T$. The probability that r's score increases at time t is given by the joint probability of a scoring event occurring at t and of r scoring it.

Letting these probabilities be independent yields

$$\Pr(\Delta s_r(t) > 0) = \Pr(\Delta s_r > 0 \,|\, \theta, \text{event}) \Pr(\text{event at } t \,|\, \theta) \ , \tag{5.1}$$

where θ parameterizes the non-ideal patterns.

Scoring events occur infrequently and independently, and are now produced by a simple non-stationary point process, in which the arrival of events varies linearly with time:

$$\Pr(\text{event at } t \,|\, \lambda_0, \alpha) = \lambda_0 + \alpha t \ . \tag{5.2}$$

The base or background rate is given by λ_0 and α parameterizes the non-stationarity, e.g., increasing ($\alpha > 0$) or decreasing ($\alpha < 0$) tempo. When $\alpha = 0$, we recover the ideal case of a Poisson process with rate λ_0.

The score of a team follows a general Bernoulli process. Given a scoring event, points are awarded to team r with some probability that is fixed for this competition, but which may vary between competitions

$$\Pr(\Delta s_r > 0 \,|\, \text{event}) = c \ , \tag{5.3}$$

and otherwise, they are awarded to team b. This scoring bias c is a probabilistic measure of r's competitive advantage over b, e.g., from a difference in skill or from exploitable features of the competition. When $c = 1/2$, we recover the ideal case of a balanced Bernoulli process, while deviations produce the more lopsided trajectories associated with non-ideal dynamics.

Across competitions with the same structure, different pairs of teams will exhibit different competitive advantages. Thus, the natural explanatory target is the distribution of the scoring imbalances $\Pr(c)$, whose natural form is a symmetric Beta distribution [18] (see *Appendix A*), the conjugate prior for the Bernoulli process. The result is a one-parameter model that quantifies the overall variability in competitive advantages across a set of competitions. The ideal case of perfectly matched teams and scoring differences due only to chance events occurs at $c = 1/2$, which is recovered in the limit of $\beta \to \infty$. Smaller values of β indicate less balanced and thus more predictable scoring dynamics across the set.

5.2.1 Predicting the winner

We supplement this parametric approach with a non-parametric measure of non-ideal behavior: the predicability of the winner from a partially unfolded competition. Having observed the first k scoring events, predicting the winning team is a kind of classification task, which we formalize as a Markov chain on the sequence of team scores (see *Appendix A*). For two-team competitions, the probability that team r wins, given current scores s_r and s_b, is

$$\begin{aligned}\Pr(r \text{ wins} \,|\, s_r, s_b) = \Pr(r \text{ wins} \,|\, s_r + 1, s_b) \cdot \hat{c} + \\ \Pr(r \text{ wins} \,|\, s_r, s_b + 1) \cdot (1 - \hat{c}) \ ,\end{aligned} \tag{5.4}$$

where $\hat{c} = s_r/(s_r + s_b)$ estimates r's competitive advantage. After each event, the classifier predicts as the winner the team with the greatest estimated odds-to-win, and its accuracy is measured by the AUC statistic [25], the probability of choosing the correct winning team.

The AUC versus k provides complete information about a competition's predictability but is not amenable to our subsequent analysis. We instead use a point measure ρ, defined as the ratio of the Markov classifier's AUC to that of an ideal competition ($c = 1/2$), when 20% of the competition has unfolded. A value of $\rho > 1$ indicates that the competition outcomes are more predictable than in the ideal case.

5.3 Competition data

Our data are drawn from the popular online game *Halo: Reach*, and span nearly 1 billion scoring events across roughly 10 million diversely structured team competitions. These competitions are divided into 125 types according to 35 structural features defining the spatial environmental, competition rules, resource quality, and whether teams had roughly equal skill (see *Appendix A*).

Halo competitions are a kind of real-time virtual combat. Human players guide their avatars through an arena containing complex terrain, coordinate actions with teammates

through visual and audio signals, and encounter opponents. A scoring event occurs when one avatar eliminates another, and this event increments the former's team score. After a short delay, the latter is returned to the competition at another arena location. Competitions end either when a fixed time limit is reached (typically 10 minutes) or when one team's score reaches some threshold (typically 50).

Only individual player skill persists across competitions. Temporary resources, whose control may yield a competitive advantage, are acquirable within a competition, e.g., highly defensible positions, high quality avatar items, and tactical information. Team membership is also temporary, being substantially randomized across competitions by the online system. These features make *Halo* competitions well suited for investigating the impact of structural heterogeneities on competition dynamics. Unlike professional sports, whose team memberships persist across competitions and which exhibit little structural variation, each *Halo* competition is roughly independent of the next, which mitigates confounding effects in characterizing the importance of structural variations.

From the scoring events within a given type of competition, we estimate both model parameters and the outcome predictability (see *Appendix A*). This produces a set of coordinates $(\hat{\lambda}_0, \hat{\alpha}, \hat{\beta}, \hat{\rho})$ and provides a compact and interpretable summary of that competition type's scoring dynamics and variability. Letting $\vec{\eta}$ denote the structural features of a given competition type, explaining variation across the estimated coordinates from variation in $\vec{\eta}$ will reveal the impact of structural features on competition dynamics, if any.

The determinants of balance β, which quantifies the strength and distribution of competitive advantages, are of particular interest. Players may prefer more balance because it offers a fair chance at winning. Or, they may prefer less balance because it offers greater reward for the risk. In these competitions, more balance moderately correlates with a lower probability that at least one player will prematurely leave the field of play ($r^2 = 0.43$, see *Appendix A*), a typically voluntary action. Thus, players exhibit a moderate but real preference for more balanced, i.e., more ideal, competitions, whose outcomes are less predictable,

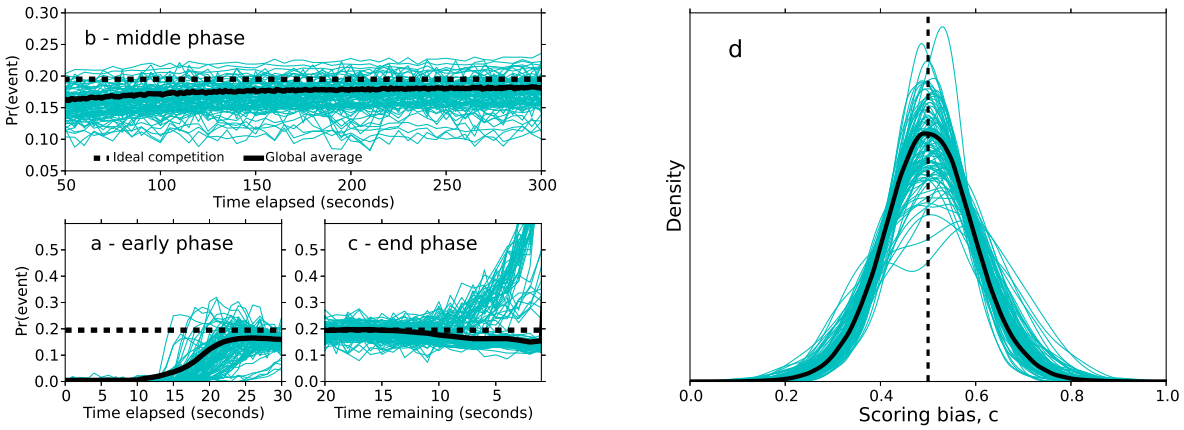

Figure 5.1: Patterns in tempo and score dynamics. For each of 125 competition types, the probability of a scoring event at time t, in the (a) early, (b) middle and (c) end phases of a competition; and (d), the distributions of the probability that team r is awarded the point. Ideal (dashed) and the global average (solid) patterns are also shown.

whose final score differences are smaller, and whose dynamics are effectively more like a simple stochastic process.

5.4 Patterns in Tempo and Score Dynamics

We first verify that our generative model effectively captures the true scoring dynamics of these competitions and whether they exhibit patterns similar to those of professional sports.

Across all competition types, we find a consistent three-phase non-stationary pattern in the tempo of scoring events, i.e., the probability of a scoring event as a function of time elapsed or time remaining. Specifically, we find an early phase of little or uneven activity, a protracted middle phase of slow and steadily increasing activity, and an end phase of either slightly decreased or markedly increased activity (Fig. 5.1 a-c).

The early- and end-phase patterns are caused by boundary effects in the length of competition, and these are also observed in professional sports [50]. Early in a competition, players require some time to move from their initial positions to their first scoring opportu-

nities, which suppresses the tempo of events relative to the ideal case. Although the shape of this early phase varies moderately by competition type (Fig. 5.1 *a*), after 20–30 seconds these variations largely disappear and the tempo transitions into the more stable middle phase.

Similarly, near a competition's end, the impending cessation of scoring opportunities encourages different strategic choices [88] than in the early or middle phases. Here, we observe either slightly decreased or strongly increased tempo (Fig. 5.1 *c*), depending on whether the competition type's particular rules provide an incentive for risk taking in the final seconds. When the incentive is present, the tempo increases dramatically just before the competition ends, as players take greater risks for the win—a pattern also observed in professional sports [50, 113]. When the incentive is absent, players instead adopt defensive positions to deny the opposing team additional points, leading to decreased scoring rates—a pattern not typically observed in sports.

In contrast, the middle phase's tempo exhibits a roughly linear increase over time (Fig. 5.1 *b*), which agrees with our generative model for event timing. To estimate our tempo model parameters, we eliminate the boundary effects by focusing on events in this phase alone (see *Appendix A*). Across competition types, both the base tempo and the acceleration vary widely: base rates can vary by up to a factor of two and we observe increases in tempo of 5–20% over the phase. Within-competition learning is one likely explanation for this increase [114]. Through trial and error, teams may learn how and where to produce scoring events, which progressively reduces the time spent searching for new scoring opportunities.

To understand the variation in the accumulation of points, we examine the distributions of scoring biases across competition types. For a particular competition, the scoring bias is estimated as the fraction of points held by an arbitrarily labeled team r. We find that all competition types exhibit moderately non-ideal variations in scoring biases (Fig. 5.1 *d*), i.e., they are consistently dispersed from the ideal case of $c = 1/2$. As with the competition tempo in the middle phase, the degree of dispersion varies substantially across competition

types, suggesting a significant role for structural variables.

As a further test of our generative model's quality for these competitions, we estimate λ_0, α and β from the entire data set, draw many synthetic competitions from the fitted model, and consider whether the simulated scoring dynamics are similar to those in the empirical data. The results indicate that the simulated competitions match the observed sequences on multiple scoring and timing statistics unrelated to parameter estimation (see *Appendix A*). This quantitative agreement indicates that our model successfully captures the important dynamical features of our competitions.

5.4.1 How structure shapes dynamics.

We now investigate four specific types of structure and their impact on the estimated competition dynamics. These analysis are intended to shed light on how specific structures may shape dynamics, and will aid the interpretation of our systematic analysis below.

Team skill differences. When assigning individuals to a new competition instance, the online system uses a matchmaking algorithm to substantially randomize team composition. This algorithm operates in two modes. For players who have completed a moderate number of competitions, it adjusts team memberships so that teams have roughly equal total skill. These estimates are derived from a Bayesian generalization of the popular Elo rating system of individual player skill [60]. Otherwise, teams are assembled without regard to player skill. We examine the differences in our model parameters for all competitions constructed under each of the two modes.

Differences in skill have a substantial impact on competition balance, as we might expect. However, they have little impact on competition tempo (Table 5.1, Fig. A.4 *a*). When teams have roughly equal skill, scoring is more balanced than when the equal-skill control is absent ($\beta = 45.9 \pm 0.35$ versus 20.9 ± 0.22). This difference implies that well-matched teams produce substantially more ideal competitions, have smaller competitive advantages, and

Table 5.1: Estimated tempo and scoring parameters for four dimensions of competition variation, illustrating a substantial impact of structure on dynamics. Values in parentheses give the bootstrap uncertainty.

feature	variation	balance $\hat{\beta}$	base tempo $\hat{\lambda}_0$ ($\times 10^{-3}$)	acceleration $\hat{\alpha}$ ($\times 10^{-5}$)
skill	equal	45.9(0.35)	166(0.1)	7.09(0.09)
	unequal	20.9(0.22)	160(0.1)	7.18(0.02)
environment	neutral	47.9(1.20)	169(0.4)	9.09(0.22)
	irregular	23.9(0.67)	147(0.3)	7.49(0.21)
scoring	standard	41.7(0.36)	185(0.2)	8.45(0.16)
	easy	30.3(0.71)	158(1.1)	9.16(0.64)
resources	versatile	20.2(0.52)	153(0.2)	7.08(0.13)
	limited	41.7(1.04)	166(0.3)	8.49(0.21)
all	–	29.5(0.21)	163(0.1)	7.13(0.05)

exhibit overall dynamics that are closer to those produced by a fair coin. In effect, reducing the difference in team skill serves to amplify the importance of chance events, i.e., accidents and miscalculations.

Physical environment. The arenas for these competitions are typically complex virtual terrains, and may contain large outdoor spaces, complicated indoor corridor systems, buildings with multiple levels, defensible positions, high ground, etc. We compare model parameters for all competitions taking place within two structurally distinct environments: one is largely neutral, exhibiting strong spatial symmetries and few features like defensible locations that might offer tactical advantage, while the other is strongly irregular, with an asymmetric and strongly vertical spatial structure, truncated sight lines, and at least one defensible location.

Overall, the more symmetric environment produces substantially more balanced outcomes and higher scoring rates than the irregular one. In fact, the observed difference in balance parameters is roughly as large as the difference induced by the equal-skill criterion (Table 5.1, Fig. A.4 *b*). This suggests that increasing the homogeneity of the competitive environment, e.g., introducing symmetries, removing defensible positions, etc., serves to limit

environmental opportunities for competitive advantage. Much like eliminating differences in skill, simpler environments effectively amplify the importance of chance events, making competition scoring more ideal.

Scoring difficulty. Few studies have examined the difference in competition dynamics caused by variations in the rules of the competition. Our data include several variations of this kind, and we examine one particular variant to shed light on how small changes in rules may impact competition dynamics. A popular group of competition types alters the standard scoring rules by reducing the threshold required to eliminate an opposing avatar and by slightly limiting each player's visual field. These changes make scoring opportunities easier to exploit, and we compare the estimated model parameters for all competitions of the standard and easy scoring types.

Lowering the threshold for scoring has a substantial impact on competition dynamics (Table 5.1, Fig. A.4 *c*), with easier scoring rules producing less balanced outcomes. The size of this difference is nearly half as large as the impact of the equal-skill criterion. Additionally, the lower threshold decreases the base scoring rate by 15% but increases the acceleration by roughly 8% over those of standard competitions. The implication is that lowering the barrier to scoring skews the playing field, allowing skilled players to exploit either their skill-based competitive advantage or other structurally-derived advantages.

Resource quality. Each competition has a fixed a set of acquirable resources, which players use to score points. Each resource belongs to one of two classes, which we label "versatile" and "limited." Versatile resources are generally of higher quality and are more effective for scoring points. When resources of both classes are present in a competition, 80% of scoring events are associated with the versatile class, illustrating a strong player preference for more effective tools. To clearly separate their effects, we examine competitions with either only versatile- or only limited-class resources.

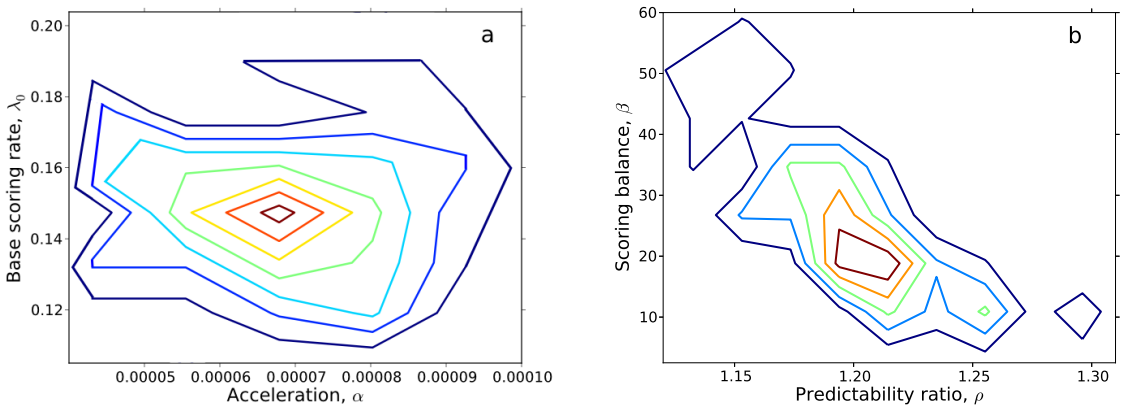

Figure 5.2: Equally spaced quantiles of joint distributions across 125 competition types of (a) base scoring rate λ_0 and acceleration α, and (b) outcome balance β and predictability ratio ρ. For event timing parameters, we observe little statistical correlation, while greater balance is strongly correlated with lower outcome predictability.

Limited-class competitions produced moderately higher base and acceleration rates than versatile-class competitions, indicating an overall faster tempo. Furthermore, competitions with only limited-class resources produce substantially more balanced scoring outcomes ($\beta = 41.7 \pm 1.04$ versus 20.2 ± 0.52; Table 5.1, Fig. A.4 d), a difference as large as that of the equal-skill criterion. Just as environmental structures can be exploited for competitive advantages, differences in the quality of acquirable resources also represent exploitable structural heterogeneities, and limiting such variations can effectively level a playing field to produce more ideal dynamics.

5.5 Structural determinants of competitive dynamics.

Each competition type defines a point on a $(\lambda_0, \alpha, \beta, \rho)$-manifold, and the distribution of these points describes the observed variability in competition dynamics. We now consider the degree to which a competition's position in this coordinate space is predictable from its structural features alone.

The joint distribution of the model timing parameters λ_0 and α is broadly distributed and shows little internal structure (Fig. 5.2a). The typical scoring base rate is roughly

one event per 7.5 seconds, with variations of 2.5s in either direction. Additionally, nearly all competitions types show modest acceleration rates, with an increase of 10–12% over the middle-phase of competition being common. The estimated balance parameters β are also broadly distributed, indicating a wide range of competitive advantages. The typical competition type has β between 20 and 30, but some have values as large as 50 or as small as 10 (Fig. 5.2*b*). We also observe a strong negative correlation between scoring balance β and the predictability ρ of a competition's winner, although with some variation, particularly in the low-β regime.

5.6 Predicting dynamics from structure

The extent to which a competition's dynamical variables $(\lambda_0, \alpha, \beta, \rho)$ are predictable from its structural variables $\vec{\eta}$ provides a direct measure of how competition structure shapes dynamics. Thirty-five structural features, divided into resources (R), environment (E), team skill (S), and rules (P) categories, were used to identify 125 distinct types of competition. Regressing these structural features onto the estimated model parameters quantifies the overall predictability of dynamics from structure. The relative importance of these features provides additional insight.

Overall, competition dynamics are highly predictable from structure alone (Table 5.2), with structural variables explaining 65–96% of the variance in individual dynamical parameters. Because the coverage across our feature space is sparse, we performed three additional tests to determine the robustness of our results. Both multiple and stepwise regressions produce models of nearly equal quality and assign features nearly the same relative importances. Randomizing the association of structural and dynamical variables yields non-significant correlations (see *Appendix A*), indicating our results are reliable.

Competition structure has the largest impact on base rate λ_0 ($r^2 = 0.96$), and features describing neutral or homogeneous environments play the dominant role in setting its value. The base scoring rate is effectively determined by the "encounter rate" between scoring

Table 5.2: Ordered multivariate regression coefficients, with uncertainty, for predicting $\hat{\beta}$, $\hat{\lambda}_0$ and $\hat{\alpha}$ of standard-type competitions from structural features alone, and the corresponding fraction of variance explained r^2. Here, we show only the statistically significant features ($p \ll 0.001$, t-test); Table S6 provides the full results.

	structural feature	$\hat{\theta}$	std. error	r^2
λ_0	E5 indoor terrain	0.082	0.008	0.96
	E11 large arena	0.059	0.003	
	E1 open terrain	0.045	0.009	
	E3 circular terrain	0.029	0.006	
	E9 outdoor terrain	0.023	0.001	
	S1 equally skilled teams	0.005	0.001	
	R1 short & medium range	-0.021	0.008	
	R4 short & long range	-0.030	0.008	
	R15 high-quality resources	-0.032	0.006	
	E2 vertical environment	-0.081	0.006	
	E7 high ground	-0.081	0.005	
α	R12 long range	$-1.9\text{x}10^{-5}$	$8.1\text{x}10^{-6}$	0.65
	S1 equally skilled teams	$-2.9\text{x}10^{-6}$	$1.7\text{x}10^{-6}$	
$\log\beta$	E5 indoor terrain	1.849	0.320	0.93
	E1 open terrain	1.391	0.371	
	E11 large arena	1.123	0.141	
	S1 equally skilled teams	0.822	0.034	
	E9 outdoor terrain	0.481	0.076	
	E6 defensible positions	-0.813	0.150	
	E2 vertical environment	-1.645	0.336	
	E7 high ground	-2.126	0.224	
ρ	E7 high ground	0.138	0.022	0.89
	E2 vertical environment	0.123	0.024	
	E6 defensible positions	0.061	0.014	
	E9 outdoor terrain	-0.036	0.007	
	S1 equally skilled teams	-0.055	0.003	
	E11 large arena	-0.089	0.013	

opportunities and competitors. In these competitions, an encounter requires two individuals to locate and engage each other; thus, small, neutral environments generate these encounters more often than large, irregular ones. Competitions between equally-skilled teams exhibit a higher encounter rate, but only marginally, as the skill coefficient is four time smaller in absolute value than any other statistically significant feature.

The change in scoring rate α is moderately well predicted by structure ($r^2 = 0.65$), and competitions with resources that operate across long ranges and with well-matched teams exhibit less acceleration over the middle phase. These resources make it easier to locate and exploit the next scoring opportunity, thus mitigating the difficulty of searching for new opportunities within large or irregular environments. Similarly, skilled competitors tend to have prior experience with the location of resources and strategic environmental structures, improving their search efficiency and lowering α.

The scoring balance β, which measures the strength of the associated competitive advantages, is highly predictable from structure ($r^2 = 0.93$, regression on $\log \beta$), as is the relative predictability ρ of the winning team ($r^2 = 0.89$). Having well-matched teams, however, is only moderately important for increasing balance, and well balanced scoring is typically derived from large, neutral environments, a situation similar to professional team sports with their level playing fields. However, the single feature that produces the most balanced competitions, by a factor of two, is indoor terrain, i.e., rooms and corridors. This particular form spatial heterogeneity may effectively handicap all competitors by limiting their spatial awareness, thus mitigating other competitive advantages, including those derived from greater skill or more versatile resources, thereby making scoring opportunities and outcomes less predictable and more ideal.

In contrast, the most imbalanced and predictable competitions are those with controllable or strategically valuable environmental features like high ground or defensible positions. For setting the values of β and ρ, such features are at least as important, but opposite in sign, to having teams of equal skill. These strategically important environmental features can thus effectively upset the competitive balance produced by well-matched teams by providing one team with a sustained competitive advantage throughout the competition.

Surprisingly, variation in rules, including reduced spatial awareness, weakened defensive capabilities, or a lower threshold for scoring, were not statistically significant predictors. None of these features produced a measurable impact on the tempo or balance of scoring

within competitions, once the effects of other features were taken into account.

5.7 Discussion

Although professional sports are often considered models of team competition [50, 113, 61, 109, 16], their limited structural variation provides few opportunities for understanding how competition structure can shape competition dynamics. Our results shed new light on these and other fundamental questions about human social dynamics and competition.

In particular, heterogeneities in the spatial environment, available resources, competition rules, and team skill exert a strong influence on the balance and tempo of scoring within a competition. For the virtual team-combat simulation studied here, spatial structure plays the most important role in producing competitive advantages, with skill and resource differences assuming supporting roles. It is thus not a superficial analogy to say that like business firms leveraging heterogeneous and scarce resources for sustained competitive advantage in a marketplace [14], teams in *Halo* leverage environmental and resource heterogeneities, like high ground and defensible positions, toward the same ends.

But unlike the pattern of either business firms or professional sports teams, some heterogeneities—in the case of *Halo*, significant indoor terrain—can effectively neutralize competitive advantages normally derived from exploitable structural features. When these "leveling" features are present, scoring outcomes are substantially more balanced than when they are absent, and this leveling effect is stronger than the one produced by having equally skilled teams. Although the precise mechanisms of these leveling effects remain unknown, their existence implies that competitive advantages are derived from specific mechanisms whose effects can be neutralized by other mechanisms. A better understanding of these mechanisms could be derived from controlled experiments with level design, and may facilitate the design of inhomogeneous competitive environments that nevertheless exhibit the balanced dynamics that homogeneous environments produce.

Otherwise, the most ideal competitions do indeed occur in large neutral spaces between

well-matched teams. It is thus no accident that professional team sports are often played in precisely this type of environment: absent spatial or resource heterogeneity, competition between skilled teams is significantly more ideal. Counterintuitively, the more ideal a competition, the more effectively it may be described as a purely random process, not despite but in fact because of the significant strategic and tactical effort behind individual events. That is, the more ideal a competition, the greater the role of chance events like miscalculations and accidents in determining the outcome. We note, however, that replacing the underlying competition mechanics by actual coin flipping seems unlikely to produce the same level or type of engagement among players and spectators.

The three-phase pattern in the tempo of events in *Halo* competitions is strikingly similar to the pattern observed in professional team sports [50]. Yet the underlying structures of most professional sports and a team combat simulation could hardly be more different. In the former, goals have fixed locations, the environment and within-competition resources are homogeneous, and teams are highly trained and persistent. In the latter, goals are highly mobile, the environment and within-competition resources are heterogeneous, and teams are largely non-persistent. The existence of a common dynamical pattern despite such differences suggests that it may be a universal feature of team competitions. The elucidation of its origin is an important open question.

Finally, we omitted explicit roles for within-team variables like team composition [57], coordination [79], and player characteristics. Their impact is implicit within the estimated model parameters, whose variation is well explained by structural variables alone. This particular result is likely supported by the substantial randomization in team membership across *Halo* competitions, which serves to mitigate any significant differences in team composition. Player and team characteristics likely play a more significant role in determining the dynamics in competitions with persistent teams or homogeneous environments, as in professional sports. A broad study of within-competition dynamics across fundamentally different types of competition may shed complementary light on the origin of competitive advantages, the

mechanisms by which specific features promote or discourage balanced outcomes, and the fundamental laws of competitive dynamics, if any.

Chapter 6

Scoring dynamics across professional team sports: tempo, balance and predictability

6.1 Introduction

Professional team sports like American football, soccer, hockey, basketball, etc. provide a rich and relatively well-controlled domain by which to study fundamental questions about the dynamics of competition [1] . In these sports, most environmental irregularities are eliminated, players are highly trained, and rules are enforced consistently. These features produce a level playing field on which competition outcomes are determined largely by a combination of skill and luck (ideally more the former than the latter).

Modern sports in particular produce large quantities of detailed data describing not only competition outcomes and team characteristics, but also the individual events within a competition, e.g., scoring events, referee calls, timeouts, ball possessions, court positions, etc. The availability of such data has enabled many quantitative analyses of individual sports [4, 16, 113, 28, 50, 55]. Relatively little work, however, has asked what patterns or principles, if any, cut across different sports, or whether there are fundamental processes governing some dynamical aspects of all such competitions. These questions are the focus of this study, and our results shed light on several other phenomena, including the roles of skill and luck in determining outcomes, and the extent to which events early in the game influence events later in the game.

[1] This chapter is under review as S. Merritt, A. Clauset, Scoring dynamics across professional team sports: tempo, balance and predictability, Journal of Quantitative Analysis of Sports, Oct. 2013

Game theory provides an attractive quantitative framework for understanding the principles and dynamics of competition [88]. Given a set of payoffs for different actions, formal game theory can identify the optimal strategy or probability distribution over actions against an intelligent adversary. In simple decision spaces, like penalty shots in soccer [98] or serve-and-return play in tennis [117], professional athletes appear to behave as game theory predicts (although some do not [100]). However, most professional team sports exhibit large and complex decision spaces, with many possible actions of uncertain payoffs, and execution is carried out by an imperfectly coordinated team. Game theory provides less guidance within such complex games, and the resulting dynamics are often better described using tools from dynamical systems [99, 51].

Using such an approach, we investigate the within-game scoring dynamics of four team sports, college and professional (American) football, professional hockey, and professional basketball. Our primary goals are (i) to quantify and identify the common empirical patterns in scoring dynamics of these sports, and (ii) to understand the competitive processes that produce these patterns. We do not consider non-stationary effects across games, e.g., evolving team rosters or skill sets, playing field variables, etc. Instead, we focus explicitly on the sequence of scoring events within games. For each sport, we study three measurable quantities: scoring event tempo, balance, and predictability. We take an inferential approach to investigating their cross-sport patterns and present a generative model of competition dynamics that can be fitted directly to scoring event data within games. We apply this model to a comprehensive data set of 1,279,901 scoring events across 9 or 10 years of consecutive seasons in our four team sports.

There are many claims in both the academic literature and the popular press about scoring dynamics within sports, and sports are often used as exemplars of decision making and dynamics in complex competitive environments [7, 12, 100, 17]. Our results on common patterns in scoring dynamics and the processes that generate them serve to clarify, and in several cases directly contradict, many of these claims, and provide a systematic perspective

on the general phenomenon.

6.1.1 Summary of results

Table 6.1 summarizes our results as they relate to a series of specific questions about scoring dynamics.

Across all sports, scoring tempo—when scoring events occur—is remarkably well-described by a Poisson process, in which scoring events occur independently with a sport-specific rate at each second on the game clock. This rate is fairly stable across the course of gameplay, except in the first and last few seconds of a scoring period, where it is much lower or much higher, respectively, than normal. This common pattern implies that scoring events are largely memoryless, i.e., the timing of events earlier in the game have little or no impact on the timing of future events. Memorylessness contrasts with the dynamics of strategic games like chess or Go, in which events early in a game constrain and drive later events. Instead, professional sports appear to exhibit little strategic entailment, and events are driven instead by short-term optimization for scoring as quickly as possible.

The scoring balance between teams—how often a team wins a scoring event—is well-described by a common Bernoulli process, with a bias parameter that varies effectively over gameplay and across sports. Football and hockey exhibit a common pattern in which the probability of scoring again while in the lead effectively increases with lead size. In basketball, however, this probability decreases with lead size (a phenomenon first identified by [50]). The former pattern is consistent with the outcome of each scoring event being determined by a memoryless coin flip whose bias depends on the difference in the teams' inherent skill levels. The pattern in basketball is also consistent with such a process, but where on-court team skill varies inversely with lead size as a result of teams deploying their weaker players when they are in the lead and their stronger players when they are not. This player management strategy produces substantially more unpredictable games than in other sports, with winning teams losing their lead and losing teams regaining it much more often

than we would normally expect.

Overall, these results reinforce the conclusions from scoring tempo, indicating that event outcomes early in a game have little or no impact on event outcomes later in the game, which reinforces statistical claims that teams do not become "hot," [116, 7, 50] with successes running in streaks. Instead, gameplay is largely a sequence of roughly independent, short-term optimizations aimed at maximizing near-term scoring rates, with little multi-play strategic efforts and few downstream consequences for mistakes or miscalculations. This memorylessness may be caused by a persistently level playing field, which lacks strategically exploitable environmental features [81] and forbids actions that might produce sustained competitive advantages [14] as a result of within-game choices, e.g., eliminating an opposing team's best players.

We combine these insights within a generative model of gameplay and demonstrate that it accurately reproduces the observed evolution of lead-sizes over the course of games in all four sports, and also makes highly accurate predictions of game outcomes, when only the first few scoring events have occurred. Cursory comparisons suggest that this model achieves accuracy comparable to or better than several commercial odds-makers, despite this model knowing nothing about teams, players, or strategies, and instead relying exclusively on the observed tempo and balance patterns in scoring events.

6.2 A null model for competition dynamics

We first introduce the limiting case of an *ideal competition*, which provides a useful tool by which to identify and quantify interesting deviations within real data, and to generate hypotheses as to what underlying processes might produce them. Although we describe this model in terms of two teams accumulating points, it can in principle be generalized to other forms of competition.

In an ideal competition, events unfold on a perfectly neutral or "level" playing field, in which there are no environmental features that could give one side a competitive advantage

Question	*Answer*
Does scoring in games of different team sports follow common patterns?	Yes. The pattern of when points are scored and who gets them are remarkably similar across sports.
What is the common pattern?	Events occur randomly (a Poisson process). Which team wins the points is coin flip (a Bernoulli process) that depends on the relative skill difference of the teams on the field.
What might cause this pattern?	A strong focus on short-term maximization of scoring opportunities, while blocking the other team from the same. There is no evidence of strategic planning across plays, as in games like chess or Go. Teams largely react to events as they occur.
What determines how often scoring occurs?	Each sport has a characteristic rate (see Table 6.3), which increases dramatically at the end of scoring periods.
What determines who wins an event?	Skill and luck, in that order.
Do events early in a game influence events later in a game?	No. Each scoring event or "play" is effectively independent, once we control for relative team skill (and lead size in basketball). Gameplay is effectively "memoryless."
Can a team be "hot," where they score in streaks?	No. Just like players [7], teams do not get "hot." Scoring streaks are caused by getting lucky.
When is it easier or harder to score?	Every moment is equally easy or difficult. But, teams try harder at the end of a period.
Which sport is the most unpredictable?	Pro basketball, where lead sizes (spreads) tend to shrink back to zero. This tendency generates many "ties" as a game unfolds.
Do other sports exhibit this pattern?	No. Pro basketball is the only sport where the spread tends to shrink. In football and hockey, the spread tends to grow over time.
Does being behind help you win, as argued by [17]?	No. Being behind helps you lose. Being ahead and being lucky helps you win.

Table 6.1: A summary of our results, in question-and-answer format.

over the other [81]. Furthermore, each side is perfectly skilled, i.e., they possess complete information both about the state of the game, e.g., the position of the ball, the location of the players, etc. and the set of possible strategies, their optimum responses, and their likelihood of being employed. This is an unrealistic assumption, as real competitors are imperfectly skilled, and possess both imperfect information and incomplete strategic knowledge of the game. However, increased skill generally implies improved performance on these characteristics, and the limiting case would be perfect skill. Finally, each side exhibits a slightly imperfect ability to execute any particular chosen strategy, which captures the fact that no side can control all variables on the field. In other words, two perfectly skilled teams competing on a level playing field will produce scoring events by chance alone, e.g., a slight miscalculation of velocity, a fumbled pass, shifting environmental variables like wind or heat, etc.

An ideal competition thus eliminates all of the environmental, player, and strategic heterogeneities that normally distinguish and limit a team. The result, particularly from the spectator's point of view, is a competition whose dynamics are fundamentally unpredictable. Such a competition would be equivalent to a simple stochastic process, in which scoring events arrive randomly, via a Poisson process with rate λ, points are awarded to each team with equal probability, as in a fair Bernoulli process with parameter $c = 1/2$, and the number of those points is an iid random variable from some sport-specific distribution.

Mathematically, let $S_r(t)$ and $S_b(t)$ denote the cumulative scores of teams r and s at time t, where $0 \leq t \leq T$ represents the game clock. (For simplicity, we do not treat overtime and instead let the game end at $t = T$.) The probability that S_r increases by k points at time t is equal to the joint probability of observing an event worth k points, scored by team r at time t. Assuming independence, this probability is

$$\Pr(\Delta S_r(t) = k) = \Pr(\text{event at } t)\Pr(r \text{ scores})\Pr(\text{points} = k) \ . \tag{6.1}$$

The evolution of the difference in these scores thus follows an finite-length unbiased random walk on the integers, moving left or right with equal probability, starting at $\Delta S = 0$ at $t = 0$.

Real competitions will deviate from this ideal because they possess various non-ideal features. The type and size of such deviations are evidence for competitive mechanisms that drive the scoring dynamics away from the ideal.

6.3 Scoring event data

Throughout our analyses, we utilize a comprehensive data set of all points scored in league games of consecutive seasons of college-level American football (NCAA Divisions 1–3, 10 seasons; 2000–2009), professional American football (NFL, 10 seasons; 2000–2009), professional hockey (NHL, 10 seasons; 2000–2009), and professional basketball (NBA, 9 seasons, 2002–2010).[2] Each scoring event includes the time at which the event occurred, the player and corresponding team that won the event, and the number of points it was worth. From these, we extract all scoring events that occurred during regulation time (i.e., we exclude all overtime events), which account for 99% or more of scoring events in each sport, and we combine events that occur at the same second of game time. Table 6.2 summarizes these data, which encompass more than 1.25 million scoring events across more than 40,000 games.

A brief overview of each sport's primary game mechanics is provided in Appendix B. In general, games in these sports are competitions between two teams of fixed size, and points are accumulated each time one team places the ball or puck in the opposing team's goal. Playing fields are flat, featureless surfaces. Gameplay is divided into three or four scoring periods within a maximum of 48 or 60 minutes (not including potential overtime). The team with the greatest score at the end of this time is declared the winner.

6.4 Game tempo

A game's "tempo" is the speed at which scoring events occur over the course of play. Past work on the timing of scoring events has largely focused on hockey and basketball [113,

[2] Data provided by STATS LLC, copyright 2013.

sport	abbrv.	seasons	teams	competitions	scoring events
Football (college)	CFB	10, 2000–2009	486	14,588	120,827
Football (pro)	NFL	10, 2000–2009	31	2,654	19,476
Hockey (pro)	NHL	10, 2000–2009	29	11,813	44,989
Basketball (pro)	NBA	9, 2002–2010	31	11,744	1,080,285

Table 6.2: Summary of data for each sport, including total number of seasons, teams, competitions, and scoring events.

50], with little work examining football or in contrasting patterns across sports. However, these studies show strong evidence that game tempo is well approximated by a homogenous Poisson process, in which scoring events occur at each moment in time independently with some small and roughly constant probability.

Analyzing the timing of scoring events across all four of our sports, we find that the Poisson process is a remarkably good model of game tempo, yielding predictions that are in good or excellent agreement with a variety of statistical measures of game play. Furthermore, these results confirm and extend previous work [7, 50] showing little or no evidence for the popular belief in "momentum" or "hot hands," in which scoring once increases the probability of scoring again very soon. However, we do find some evidence for modest non-Poissonian patterns in tempo, some of which are common to all four sports.

6.4.1 The Poisson model of tempo

A Poisson process is fully characterized by a single parameter λ, representing the probability that an event occurs, or the expected number of events, per unit time. In each sport, game time is divided into seconds and there are T seconds per a game (see Table 6.3). For each sport, we test this model in several ways: we compare the empirical and predicted distributions for the number of events per game and for the time between consecutive scoring events, and we examine the two-point correlation function for these inter-event times.

Under a Poisson model [20], the number of scoring events per game follows a Poisson

sport	$\hat{\lambda}$ [events / s]	T [s]	$\hat{\lambda}T$ [events / game]	$1/\hat{\lambda}$ [s / event]
NFL	0.00204(1)	3600	7.34	490.2
CFB	0.00230(1)	3600	8.28	434.8
NHL	0.00106(1)	3600	3.81	943.4
NBA	0.03194(5)	2880	91.99	31.3

Table 6.3: Tempo summary statistics for each sport, along with simple derived values for the expected number of events per game and seconds between events. Parenthetical values indicate standard uncertainty in the final digit.

distribution with parameter λT, and the maximum likelihood estimate of λ is the average number of events observed in a game divided by the number of intervals (which varies per sport). Furthermore, the time between consecutive events follows a simple geometric (discrete exponential) distribution, with mean $1/\lambda$, and the two-point correlation between these delays is zero at all time scales.

For the number of events per game, we find generally excellent agreement between the Poisson model and the data for every sport (Figure 6.1). However, there are some small deviations, which suggests some second-order, non-Poissonian processes, which we investigate below. Deviations are greatest in NHL games, whose distribution is slightly broader than predicted, underproducing games with 3 events, and overproducing games with 0 or with 8 or more events. Similarly, CFB games have a slight excess of games with 9 events, and NBA games exhibit slightly more variation in NBA games with scores close to the average (92.0 events) than expected. In contrast, NFL games exhibit slightly less variance than expected, with more games close to the average (7.3 events) than expected.

For the time between consecutive scoring events within a game, or the inter-arrival time distribution, we again find excellent agreement between the Poisson model and the data in all sports (Figure 6.2). That being said, in CFB, NFL and NBA games, there are slightly fewer gaps of the minimum size than predicted by the model. This indicates a slight dispersive effect in the timing of events, perhaps caused by the time required to transport the

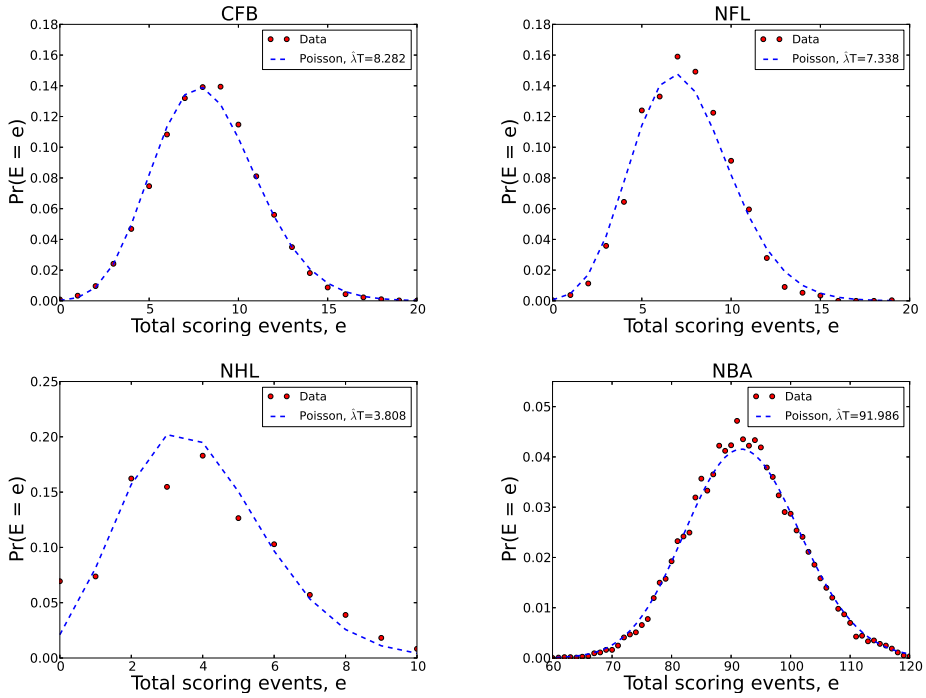

Figure 6.1: Empirical distributions for the number of scoring events per game, along with the estimated Poisson model with rate λT (dashed).

ball some distance before a new event may be generated. In contrast, NHL games produce as many short gaps, more intermediate gaps, and fewer very long gaps than expected were events purely Poissonian.

Finally, we calculate the two-point correlation function on the times between scoring events [24],

$$C(n) = \left(\sum_k (t_k - \langle t \rangle)(t_{k+n} - \langle t \rangle) \right) \Bigg/ \sum_k (t_k - \langle t \rangle)^2 \quad , \tag{6.2}$$

where t_k is the kth inter-arrival time, n indicates the gap between it and a subsequent event, and $\langle t \rangle$ is the mean time between events. If $C(n)$ is positive, short intervals tend to be followed by other short intervals (or, large intervals by large intervals), while a negative value implies alternation, with short intervals followed by long, or vice versa. Across all four sports, the correlation function is close or very close to zero for all values of n (Figure 6.2 insets), in excellent agreement with the Poisson process, which predicts $C(n) = 0$ for all $n > 0$, representing no correlation in the timing of events (a result also found by [50] in

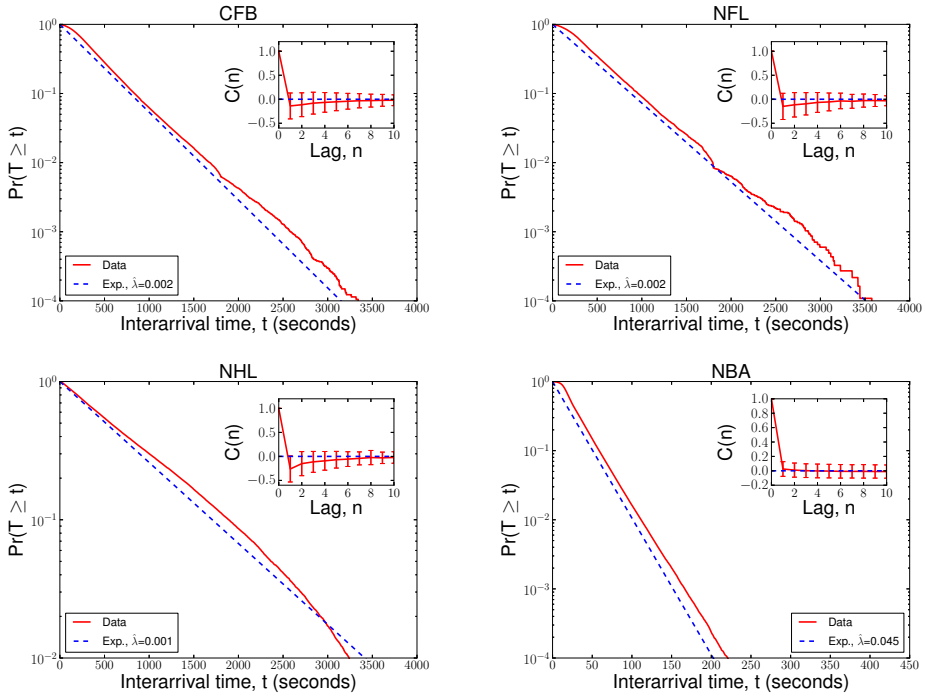

Figure 6.2: Empirical distribution of time between consecutive scoring events, shown as the complementary cdf, along with the estimated distribution from the Poisson model (dashed). Insets show the correlation function for inter-event times.

basketball). However, in CFB, NFL and NHL games, we find a slight negative correlation for very small values of n, suggesting a slight tendency for short intervals to be closely followed by longer ones, and vice versa.

6.4.2 Common patterns in game tempo

Our results above provide strong support for a common Poisson process as an excellent explanatory model of game tempo across all four sports. We also find some evidence for mild non-Poissonian processes, which we now investigate by directly examining the scoring rate as a function of clock time. Within each sport, we tabulate the fraction of games in which a scoring event (associated with any number of points) occurred in the tth second of gameplay.

Across all sports, we find that the tempo of events follows a common three-phase

pattern within each distinct period of play (Figure 6.3). This pattern, which resembles an inverse sigmoid, is characterized by (i) an early phase of non-linearly increasing tempo, (ii) a middle phase of stable (Poissonian) or slightly increasing tempo, and (iii) an end phase of sharply increasing tempo. This pattern is also observed in certain online games [81], which have substantially different rules and are played in highly heterogeneous environments, suggesting a possibly fundamental generating mechanism for team-competitive systems.

Early phase: non-linear increase in tempo. When a period begins, players are in specific and fixed locations on the field, and the ball or puck is far from any team's goal. Thus, without regard to other aspects of the game, it must take some time for players to move out of these initial positions and to establish scoring opportunities. This would reduce the probability of scoring relative to the game average by limiting access to certain player-ball configurations that require time to set up. Furthermore, and potentially most strongly in the first of these phases (beginning at $t = 0$), players and teams may still be "warming up," in the sense of learning [114] the capabilities and tendencies of the opposing team and players, and which tactics to deploy against the opposing team's choices. These behaviors would also reduce the probability of scoring by encouraging risk averse behavior in establishing and taking scoring opportunities.

We find evidence for both mechanisms in our data. Both CFB and NFL games exhibit short and modest-sized dips in scoring rates in periods 2 and 4, reflecting the fact that player and ball positions are not reset when the preceding quarters end, but rather gameplay in the new quarter resumes from its previous configuration. In contrast, CFB and NFL periods 1 and 3 show significant drops in scoring rates, and both of these quarters begin with a kickoff from fixed positions on the field. Similarly, NBA and NHL games exhibit strong but short-duration dips in scoring rate at the beginning of each of their periods, reflecting the fact that each quarter begins with a tossup or face-off, in which players are located in fixed positions on the court or rink. NBA and football games also exhibit some evidence of the "warming up" process, with the overall scoring rate being slightly lower in period

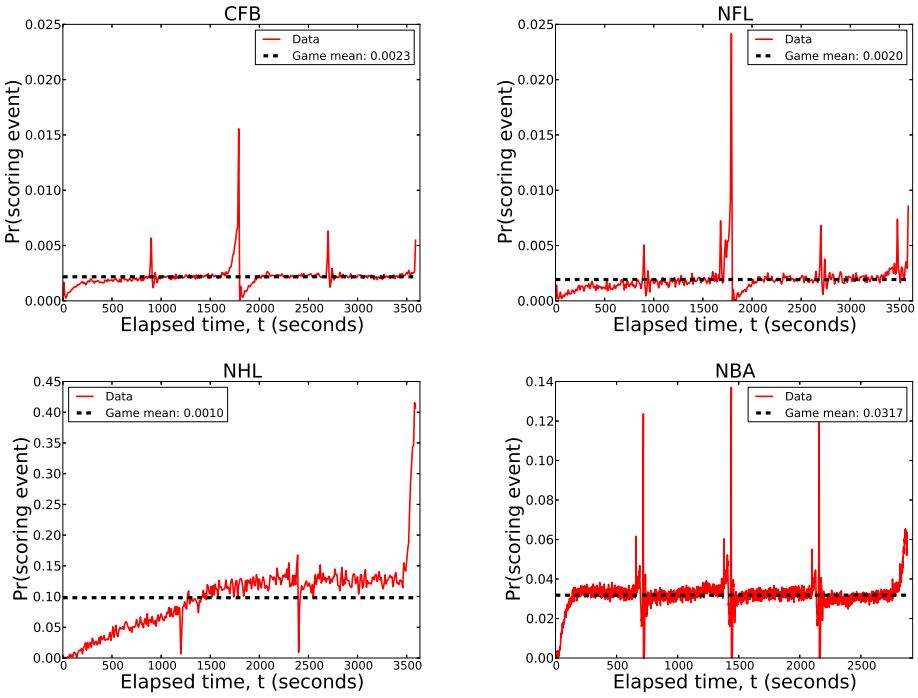

Figure 6.3: Empirical probability of scoring events as a function of game time, for each sport, along with the mean within-sport probability (dashed line). Each distinct game period shows a common three-phase pattern in tempo.

1 than in other equivalent periods. In contrast, NHL games exhibit a prolonged warmup period, lasting well past the end of the first period. This pattern may indicate more gradual within-game learning in hockey, perhaps are a result of the large diversity of on-ice player combinations caused by teams rotating their four "lines" of players every few minutes.

Middle phase: constant tempo. Once players have moved away from their initial locations and/or warmed up, gameplay proceeds fluidly, with scoring events occurring without any systematic dependence on the game clock. This produces a flat, stable or stationary pattern in the probability of scoring events. A slight but steady increase in tempo over the course of this phase is consistent with learning, perhaps as continued play sheds more light on the opposing team's capabilities and weaknesses, causing a progressive increase in scoring rate as that knowledge is accumulated and put into practice.

A stable scoring rate pattern appears in every period in NFL, CFB and NBA games,

with slight increases observed in periods 1 and 2 in football, and in periods 2–4 in basketball. NHL games exhibit stable scoring rates in the second half of period 2 and throughout period 3. Within a given game, but across scoring periods, scoring rates are remarkably similar, suggesting little or no variation in overall strategies across the periods of gameplay.

End phase: sharply increased tempo. The end of a scoring period often requires players to reset their positions, and any effort spent establishing an advantageous player configuration is lost unless that play produces a scoring event. This impending loss-of-position will tend to encourage more risky actions, which serve to dramatically increase the scoring rate just before the period ends. The increase in scoring rate should be largest in the final period, when no additional scoring opportunities lay in the future. In some sports, teams may effectively slow the rate by which time progresses through game clock management (e.g., using timeouts) or through continuing play (at the end of quarters in football). This effectively compresses more actions than normal into a short period of time, which may also increase the rate, without necessarily adding more risk.

We find evidence mainly for the loss-of-position mechanism, but the rules of these games suggest that clock management likely also plays a role. Relative to the mean tempo, we find a sharply increased rate at the end of each sport's games, in agreement with a strong incentive to score before a period ends. (This increase indicates that a "lolly-gag strategy," in which a leading team in possession intentionally runs down the clock to prevent the trailing team from gaining possession, is a relatively rare occurrence.) Intermediate periods in NFL, CFB and NBA games also exhibit increased scoring rates in their final seconds. In football, this increase is greatest at the end of period 2, rather than period 4. The increased rate at the ends of periods 1 and 3 in football is also interesting, as here the period's end does not reset the player configuration on the field, but rather teams switch goals. This likely creates a mild incentive to initiate some play before the period ends (which is allowed to finish, even if the game clock runs out). NHL games exhibit no discernible end-phase pattern in their intermediate periods (1 and 2), but show an enormous end-game effect, with the scoring rate

growing to more than three times its game mean. This strong pattern may be related to the strategy in hockey of the losing team "pulling the goalie," in which the goalie leaves their defensive position in order to increase the chances of scoring. Regardless of the particular mechanism, the end-phase pattern is ubiquitous.

In general, we find a common set of modest non-Poissonian deviations in game tempo across all four sports, although the vast majority of tempo dynamics continue to agree with a simple Poisson model.

6.5 Game balance

A game's "balance" is the relative distribution of scoring events (not points) among the teams. Perfectly balanced games, however, do not always result in a tie. In our model of competition, each scoring event is awarded to one team or the other by a Bernoulli process, and in the case of perfect balance, the probability is equal, at $c = 1/2$. The expected fraction of scoring events won by a team is also $c = 1/2$, and its distribution depends on the number of scoring events in the game. We estimate this null distribution by simulating perfectly balanced games for each sport, given the empirical distribution of scoring events per game (see Fig. 6.1). Comparing the simulated distribution against the empirical distribution of c provides a measure of the true imbalance among teams, while controlling for the stochastic effects of events within games.

Across all four sports, we find significant deviations in this fraction relative to perfect balance. NFL and CFB games exhibited more variance than expected, while NHL and NBA games exhibited the least. Within a game, scoring balance exhibits unexpected patterns. In particular NBA games exhibit an unusual "restoring force" pattern, in which the probability of winning the next scoring event *decreases* with the size of a team's lead (a pattern first observed by [50]). In contrast, NFL, CFB and NHL games exhibit the opposite effect, in which the probability of winning the next scoring event appears to increase with the size of the lead—a pattern consistent with a heterogeneous distribution of team skill.

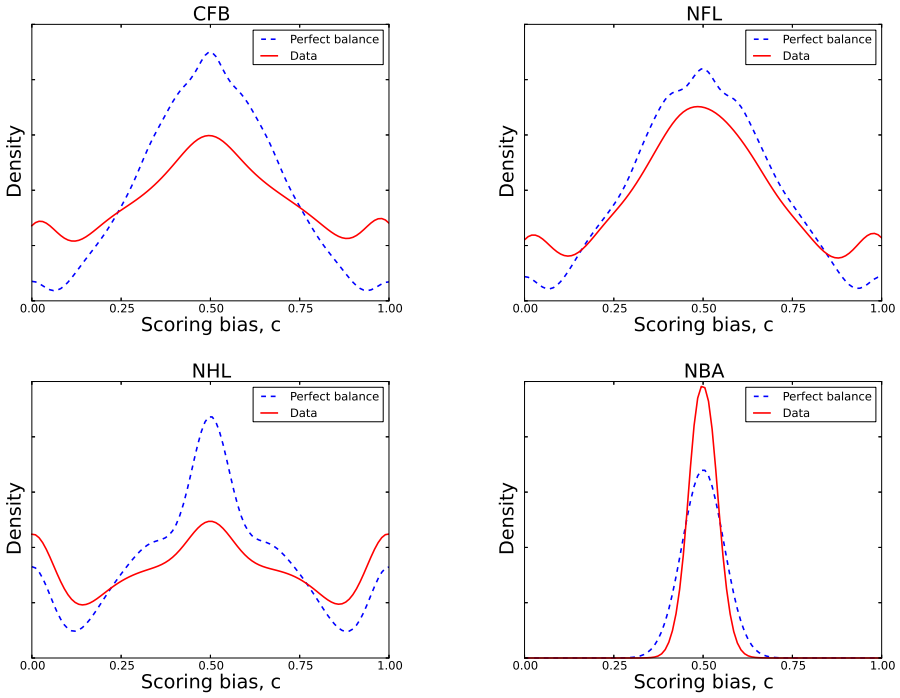

Figure 6.4: Smoothed distributions for the empirical fraction $\hat{c}$ of events won by a team, for each sport, and the predicted fraction for a perfectly balanced scoring, when given the empirical distribution of events per game (Fig. 6.1). Modes at 1 and 0 indicate a non-trivial probability of one team winning or losing every event, which is more common when only a few events occur.

6.5.1 Quantifying balance

The fraction of all events in the game that were won by a randomly selected team provides a simple measure of the overall balance of a particular game in a sport. Let r and b index the two teams and let E_r (E_b) denote the total number of events won by team r in its game with b. The maximum likelihood estimator for a game's bias is simply the fraction $\hat{c} = E_r/(E_r + E_b)$ of all scoring events in the game won by r.

Tabulating the empirical distributions of $\hat{c}$ within each sport, we find that the most common outcome, in all sports, is $c = 1/2$, in agreement with the Bernoulli model. However, the distributions around this value deviate substantially from the form expected for perfect balance (Figure 6.4), but not always in the same direction.

In CFB and NFL, the distributions of scoring balances are similar, but the shape for

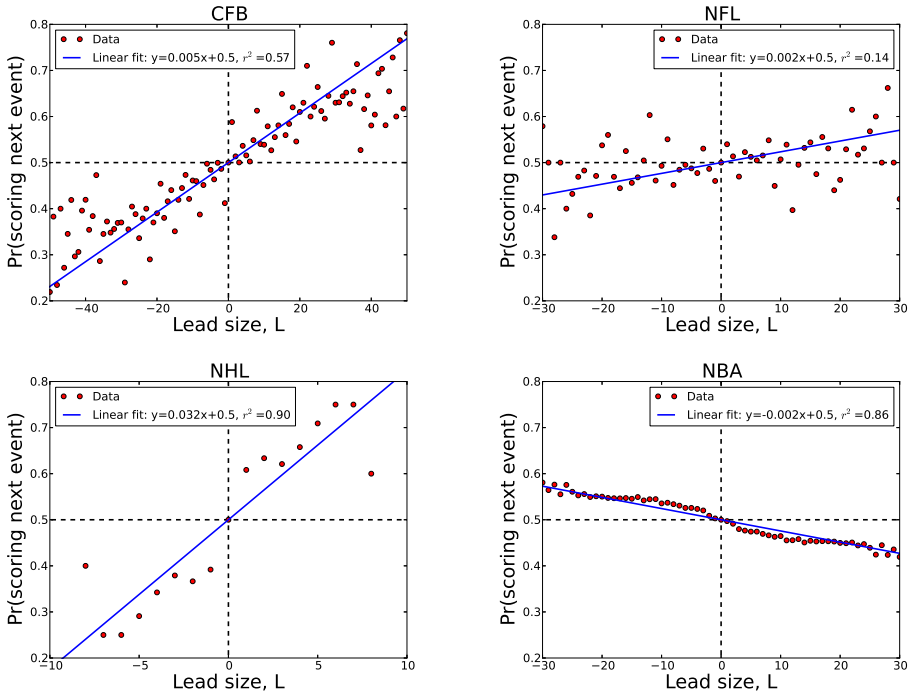

Figure 6.5: The probability of scoring as a function of a team's lead size for each sport, football, hockey, and basketball and a linear least-squares fit ($p \leq 0.1$), indicating positive or negative correlations between scoring and a competition's score difference.

CFB is broader than for NFL, suggesting that CFB competitions are less balanced than NFL competitions. This is likely a result of the broader range of skill differences among teams at the college level, as compared to the professionals. Like CFB and NFL, NHL games also exhibit substantially more blowouts and fewer ties than expected, which is consistent with a heterogeneous distribution of team skills. Surprisingly, however, NBA games exhibit less variance in the final relative lead size than we expect for perfectly balanced games, a pattern we will revisit in the following section.

6.5.2 Scoring while in the lead

Although many non-Bernoulli processes may occur within professional team sports, here we examine only one: whether the size of a lead L, the difference in team scores or point totals, provides information about the probability of a team winning the next event.

[50] previously considered this question for scoring events and lead sizes within NBA games, but not other sports. Across all four of our sports, we tabulated the fraction of times the leading team won the next scoring event, given it held a lead of size L. This function is symmetric about $L = 0$, where it passes through probability $p = 1/2$ where the identity of the leading team may change.

Examining the empirical scoring functions (Figure 6.5), we find that the probability of scoring next varies systematically with lead size L. In particular, for CFB, NFL and NHL games, the probability appears to increase with lead size, while it decreases in NBA games. The effect of the negative relationship in NBA games is a kind of "restoring force," such that leads of any size tend to shrink back toward a tied score. This produces a narrower distribution of final lead sizes than we would expect under Bernoulli-style competition, precisely as shown in Figure 6.4 for NBA games.

Although the positive function for CFB, NFL and NHL games may superficially support a kind of "hot hands" or cumulative advantage-type mechanism, in which lead size tends to grow superlinearly over time, we do not believe this explains the observed pattern. A more plausible mechanism is a simple heterogeneous skill model, in which each team has a latent skill value π_r, and the probability that team r wins a scoring event against b is determined by a Bernoulli process with $c = \pi_r / (\pi_r + \pi_b)$. (This model is identical to the popular Bradley-Terry model of win-loss records of teams [26], except here we apply it to each scoring event within a game.)

For a broad class of team-skill distributions, this model produces a scoring function with the same sigmoidal shape seen here, and the linear pattern at $L = 0$ is the result of averaging over the distribution of biases c induced by the team skill distribution. The function flattens out at large $|L|$ assuming the value representing the largest skill difference possible among the league teams. This explanation is supported by the stronger correlation in CFB games (+0.005 probability per point in the lead) versus NFL games (+0.002 probability per point), as CFB teams are known to exhibit much broader skill differences than NFL teams, in

agreement with our results above in Figure 6.4.

NBA games, however, present a puzzle, because no distribution of skill differences can produce a negative correlation under this latent-skill model. [50] suggested this negative pattern could be produced by possession of the ball changing after each scoring event, or by the leading team "coasting" and thereby playing below their true skill level. However, the change-of-possession rule also exists in CFB and NFL games (play resumes with a faceoff in NHL games), but only NBA games exhibit the negative correlation. Coasting could occur for psychological reasons, in which losing teams play harder, and leading teams less hard, as suggested by [17]. Again, however, the absence of this pattern in other sports suggest that the mechanism is not psychological.

A plausible alternative explanation is that NBA teams employ various strategies that serve to change the ratio $c = \pi_r / (\pi_r + \pi_b)$ as a function of lead size. For instance, when a team is in the lead, they often substitute out their stronger and more offensive players, e.g., to allow them to rest or avoid injury, or to manage floor spacing or skill combinations. When a team is down by an amount that likely varies across teams, these players are put back on the court. If both teams pursue such strategies, then effective ratio c will vary inversely with lead size such that the leading team becomes effectively weaker compared to the non-leading team. In contrast to NBA teams, teams in CFB, NFL and NHL seem less able to pursue such a strategy. In football, substitutions are relatively uncommon, implying that π_r should not vary much over the course of a game. In hockey, each team rotates through most of its players every few minutes, which limits the ability for high- or low-skilled players to effectively change π_r over the course of a game.

6.6 Modeling lead-size dynamics

The previous insights identify several basic patterns in scoring tempo and balance across sports. However, we still lack a clear understanding of the degree to which any of these patterns is necessary to produce realistic scoring dynamics. Here, we investigate this

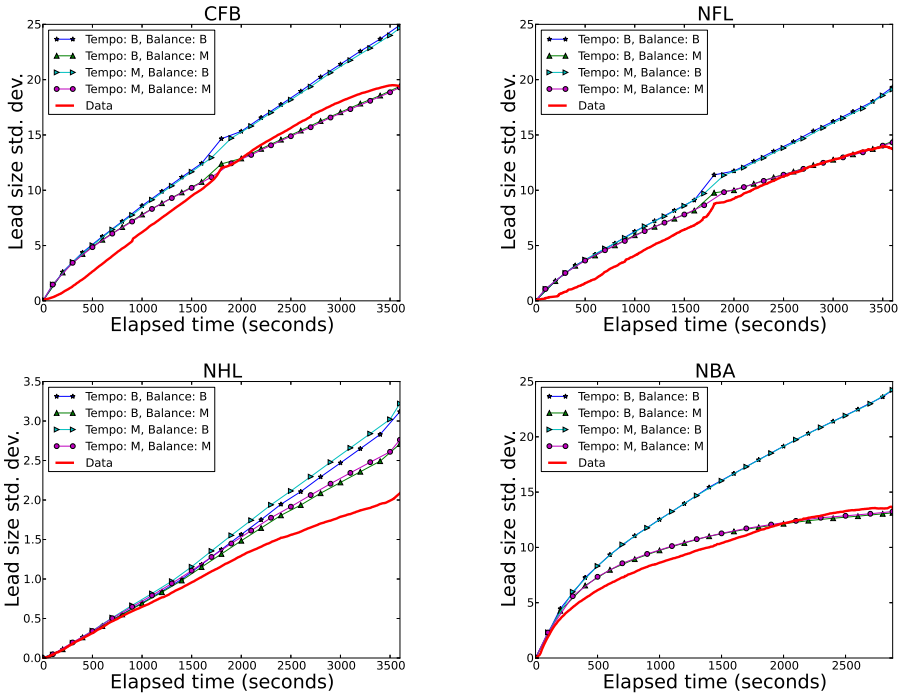

Figure 6.6: Comparison of empirical lead-size variation as a function of clock time with those produced by Bernoulli (B) or Markov (M) tempo or balance models, for each sport.

question by combining the identified patterns within a generative model of scoring over time, and test which combinations produce realistic dynamics in lead sizes. In particular, we consider two models of tempo and two models of balance. For each of the four pairs of tempo and balance models for each sport, we generate via Monte Carlo a large number of games and measure the resulting variation in lead size as a function of the game clock, which we then compare to the empirical pattern.[3]

Our two scoring tempo models are as follows. In the first (Bernoulli) model, each second of time produces an event with the empirical probability observed for that second across all games (shown in Figure 6.3). In the second (Markov), we draw an inter-arrival time from the empirical distribution of such gaps (shown in Figure 6.2), advance the game clock that that amount, and generate a scoring event at that clock time.

Our two balance models are as follows. In the first (Bernoulli) model, for each match

[3] Our simulation code may be found at `https://github.com/merritts/SportsSciPy`.

we draw a uniformly random value c from the empirical distribution of scoring balances (shown in Figure 6.4) and for each scoring event, the points are won by team r with that probability and by team b otherwise. In the second (Markov), a scoring event is awarded to the leading team with the empirically estimated probability for the current lead size L (shown in Figure 6.5). Once a scoring event is generated and assigned, that team's score is incremented by a point value drawn iid from the empirical distribution of point values per scoring event for the sport (see Appendix B).

The four combinations of tempo and balance models thus cover our empirical findings for patterns in the scoring dynamics of these sports. The simpler models (called Bernoulli) represent dynamics with little or no memory, in which each event is an iid random variable, albeit drawn from a data-driven distribution. The more complicated models (called Markov) represent dynamics with some degree of memory, allowing past events to influence the ongoing gameplay dynamics.

Generating 100,000 competitions under each combination of models for each sport, we find a consistent pattern across sports (Figure 6.6): the Markov model of game tempo provides little improvement over the Bernoulli model in capturing the empirical pattern of lead-size variation, while the Markov model for balance provides a significant improvement over the Bernoulli model. In particular, the Markov model generates gameplay dynamics in very good agreement with the empirical patterns.

That being said, some small deviations remain. For instance, the Markov model slightly overestimates the lead-size variation in the first half, and slightly underestimates it in the second half of CFB games. In NFL games, it provides a slight overestimate in first half, but then converges on the empirical pattern in the second half. NHL games exhibit the largest and most systematic deviation, with the Markov model producing more variation than observed, particularly in the game's second half. However, it should be noted that the low-scoring nature of NHL means that what appears to be a visually large overestimate here (Fig. 6.6) is small when compared to the deviations seen in the other sports. NBA

games exhibit a similar pattern to CFB games, but the crossover point occurs at the end of period 3, rather than at period 2. These modest deviations suggest the presence of still other non-ideal processes governing the scoring dynamics, particularly in NHL games.

We emphasize that the Markov model's accuracy for CFB, NFL and NHL games does not imply that individual matches follow this pattern of favoring the leader. Instead, the pattern provides a compact and efficient summary of scoring dynamics conditioned on unobserved characteristics like team skill. Our model generates competition between two featureless teams, and the Markov model provides a data-driven mechanism by which some pairs of teams may behave as if they have small or large differences in latent skill. It remains an interesting direction for future work to investigate precisely how player and team characteristics determine team skill, and how team skill impacts scoring dynamics.

6.7 Predicting outcomes from gameplay

The accuracy of our generative model in the previous section suggest that it may also produce accurate predictions of the game's overall outcome, after observing only the events in the first t seconds of the game. In this section, we study the predictability of game outcome using the Markov model for scoring balance, and compare its accuracy to the simple heuristic of guessing the winner to be the team currently in the lead at time t. Thus, we convert our Markov model into an explicit Markov chain on the lead size L, which allows us to simulate the remaining $T - t$ seconds conditioned on the lead size at time t. For concreteness, we define the lead size L relative to team r, such that $L < 0$ implies that b is in the lead.

The Markov chain's state space is the set of all possible lead sizes (score differences between teams r and b), and its transition matrix P gives the probability that a scoring event changes a lead of size L to one of size L'. If r wins the event, then $L' = L + k$, where k is the event's point value, while if b wins the event, then $L' = L - k$. Assuming the value

and winner of the event are independent, the transition probabilities are given by

$$P_{L,L+k} = \Pr(r \text{ scores} \mid L) \Pr(\text{point value} = k)$$
$$P_{L,L-k} = (1 - \Pr(r \text{ scores} \mid L)) \Pr(\text{point value} = k) \ ,$$

where, for the particular sport, we use the empirical probability function for scoring as a function of lead size (Figure 6.5), from r's perspective, and the empirical distribution (Appendix B) for the point value.

The probability that team r is the predicted winner depends on the probability distribution over lead sizes at time T. Because scoring events are conditionally independent, this distribution is given by P^n, where n is the expected number of scoring events in the remaining clock time $T - t$, multiplied by a vector S_0 representing the initial state $L = 0$. Given a choice of time t, we estimate $n = \sum_{w=t}^{T} \Pr(\text{event} \mid w)$, which is the expected number of events given the empirical tempo function (Fig. 6.3, also the Bernoulli tempo model in Section 6.6) and the remaining clock time. We then convert this distribution, which we calculate numerically, into a prediction by summing probabilities for each of three outcomes: r wins (states $L > 0$), r ties b, (state $L = 0$), and b wins (states $L > 0$). In this way, we capture the information contained in magnitude of the current lead, which is lost when we simply predict that the current leader will win, regardless of lead size.

We test the accuracy of the Markov chain using an out-of-sample prediction scheme, in which we divide each sports' game data into a training set of a randomly selected 3/4 of all games and a test set of the remaining 1/4. From each training set, we estimate the empirical functions used in the model and compute the Markov chain's transition matrix. Then, across the games in each test set, we measure the fraction of times the Markov chain's prediction is correct. This fraction is equivalent to the popular AUC statistic [25], where AUC $= 0.5$ denotes an accuracy no better than guessing.

Instead of evaluating the model at some arbitrarily selected time, we investigate how outcome predictability evolves over time. Specifically, we compute the AUC as a function of

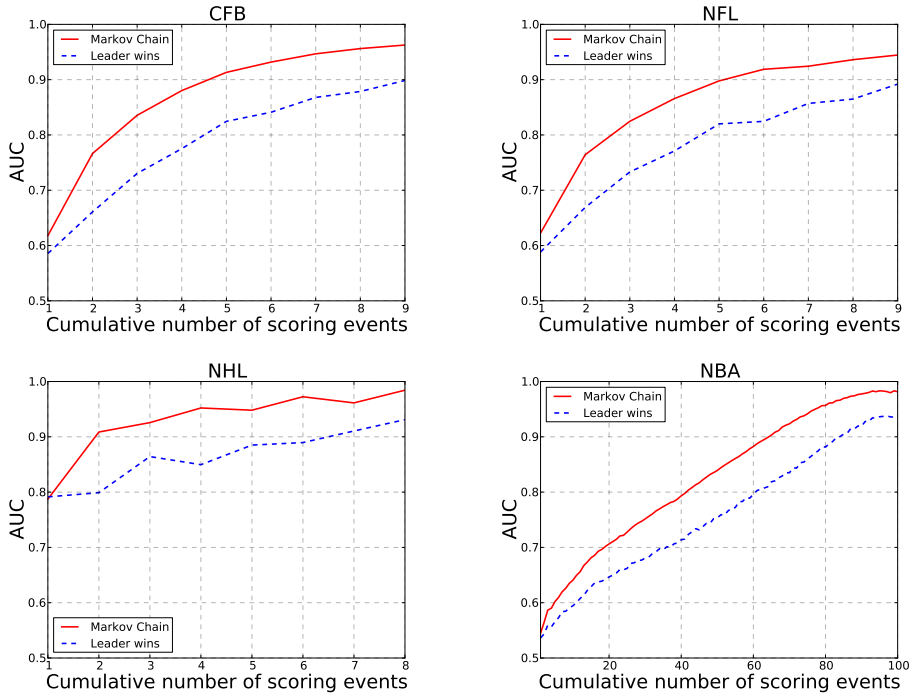

Figure 6.7: AUC's of the Markov chain and leader wins predictions of game winner for each sport, football, hockey, and basketball.

the cumulative number of scoring events in the game, using the empirically observed times and lead sizes in each test-set game to parameterize the model's predictions. When the number of cumulative events is small, game outcomes should be relatively unpredictable, and as the clock runs down, predictability should increase. To provide a reference point for the quality of these results, we also measure the AUC over time for a simple heuristic of predicting the winner as the team in the lead after the event.

Across all sports, we find that game outcome is highly predictable, even after only a small number of scoring events (Figure 6.7). For instance, the winner of CFB and NFL games can be accurately chosen more than 60% of the time after only a single scoring event, and this rate increases to more than 80% by three events. NHL games are even more predictable, in part because they are very low-scoring games, and the winner may be accurately chosen roughly 80% of the time after the first event. The fast rise of the AUC curve as a function of continued scoring in these sports likely reflects the role played by differences in latent

team skill in producing large leads, which make outcomes more predictable (Figure 6.5). In contrast, NBA games are the least predictable, requiring more than 40 events before the AUC exceeds 80%. This likely reflects the role of the "restoring force" (Figure 6.5), which tends to make NBA games more unpredictable than we would expect from a simple model of scoring, and significantly more unpredictable than CFB, NFL or NHL games.

In all cases, the Markov chain substantially outperforms the "leader wins" heuristic, even in the low-scoring NHL games. This occurs in part because small leads are less informative than large leads for guessing the winner, and the heuristic does not distinguish between these.

6.8 Discussion

Although there is increasing interest in quantitative analysis and modeling in sports [5, 67, 39, 49, 93], many questions remain about what patterns or principles, if any, cut across different sports, what basic dynamical processes provide good models of within-game events, and the degree to which the outcomes of games may be predicted from within-game events alone. The comprehensive database of scoring events we use here to investigate such questions is unusual for both its scope (every league game over 9–10 seasons), its breadth (covering four sports), and its depth (timing and attribution information on every point in every game). As such, it offers a number of new opportunities to study competition in general, and sports in particular.

Across college (American) football (CFB), professional (American) football (NFL), professional hockey (NHL) and professional basketball (NBA) games, we find a number of common patterns in both the tempo and balance of scoring events. First, the timing of events in all four sports is remarkably well-modeled by a simple Poisson process (Figures 6.1 and 6.2), in which each second of gameplay produces a scoring event independently, with a probability that varies only modestly over the course of a game (Figure 6.3). These variations, however, follow a common three-phase pattern, in which a relatively constant

rate is depressed at the beginning of a scoring period, and increases dramatically in the final few seconds of the period. The excellent agreement with a Poisson process implies that teams employ very few strategically-chosen chains of events or time-sensitive strategies in these games, except in a period's end-phase, when the incentive to score in elevated. These results provide further support to past analyses [7, 50] showing no evidence for the popular notion of "hot hands," in which scoring once increases the chance of scoring again soon.

Second, we find a common pattern of imbalanced scoring between teams in CFB, NFL and NHL games, relative to an ideal model in which teams are equally likely to win each scoring event (Figure 6.4). CFB games are much less balanced than NFL games, suggesting that the transition from college to professional tends to reduce the team skill differences that generate lopsided scoring. Furthermore, we find that all three of these sports exhibit a pattern in which lead sizes tend to increase over time. That is, the probability of scoring while in the lead tends to be larger the greater the lead size (Figure 6.5), in contrast to the ideal model in which lead sizes increase or decrease with equal probability. As with overall scoring balance, the size of this effect in CFB games is much larger (about 2.5 times larger) than in NFL games, and is consistent with a reduction in the variance of the distribution of skill across teams. That is, NFL teams are generally closer in team skill than CFB teams, and this produces gameplay that is much less predictable. Both of these patterns are consistent with a kind of Bradley-Terry-type model in which each scoring event is a contest between the teams.

NBA games, however, present the opposite pattern: team scores are much closer than we would expect from the ideal model, and the probability of scoring while in the lead effectively *decreases* as the lead size grows (Figure 6.5; a pattern originally identified by [50]). This pattern produces a kind of "restoring force" such that leads tend to shrink until they turn into ties, producing games that are substantially more unpredictable. Unlike the pattern in CFB, NFL and NHL, no distribution of latent team skills, under a Bradley-Terry-type model, can produce this kind of negative correlation between the probability of scoring and

lead size.

Recently, [17] analyzed similar NBA game data and argued that increased psychological motivation drives teams that are slightly behind (e.g., by one point at halftime) to win the game more often than not. That is, losing slightly is good for winning. Our analysis places this claim in a broader, more nuanced context. The effective restoring force is superficially consistent with the belief that losing in NBA games is "good" for the team, as losing does indeed empirically increase the probability of scoring. However, we find no such effect in CFB, NFL or NHL games (Figure 6.5), suggesting either that NBA players are more poorly motivated than players in other team sports or that some other mechanism explains the pattern.

One such mechanism is for NBA teams to employ strategies associated with substituting weaker players for stronger ones when they hold various leads, e.g., to allow their best players to rest or avoid injury, manage floor spacing and offensive/defensive combinations, etc., and then reverse the process when the other team leads. In this way, a team will play more weakly when it leads, and more strongly when it is losing, because of personnel changes alone rather than changes in morale or effort. If teams have different thresholds for making such substitutions, and differently skilled best players, the averaging across these differences would produce the smooth pattern observed in the data. Such substitutions are indeed common in basketball games, while football and hockey teams are inherently less able to alter their effective team skill through such player management, which may explain the restoring force's presence in NBA games and its absence in CFB, NFL or NHL games. It would be interesting to determine whether college basketball games exhibit the same restoring force, and the personnel management hypothesis could be tested by estimating the on-court team's skill as a function of lead size.

The observed patterns we find in the probability of scoring while in the lead are surprisingly accurate at reproducing the observed variation in lead-size dynamics in these sports (Figure 6.6), and suggest that this one pattern provides a compact and mostly accurate

summary of the within-game scoring dynamics of a sport. However, we do not believe these patterns indicate the presence of any feedbacks, e.g., "momentum" or cumulative advantage [40]. Instead, for CFB, NFL and NHL games, this pattern represents the the distribution of latent team skills, while for NBA games, it represents strategic decisions about which players are on the court as a function of lead size.

This pattern also makes remarkably good predictions about the overall outcome of games, even when given information about only the first ℓ scoring events. Under a controlled out-of-sample test, we found that CFB, NFL and NHL games are highly predictable, even after only a few events. In contrast, NBA games were significantly less predictable, although reasonable predictions here can still be made, despite the impact of the restoring force.

Given the popularity of betting on sports, it is an interesting question as to whether our model produces better or worse predictions than those of established odds-makers. To explore this question, we compared our model against two such systems, the online live-betting website Bovada[4] and the odds-maker website Sports Book Review (SBR).[5] Neither site provided comprehensive coverage or systematic access, and so our comparison was necessarily limited to a small sample of games. Among these, however, our predictions were very close to those of Bovada, and, after 20% of each game's events had occurred, were roughly 10% more accurate than SBR's money lines across all sports. Although the precise details are unknown for how these commercial odds were set, it seems likely that they rely on many details omitted by our model, such as player statistics, team histories, team strategies and strengths, etc. In contrast, our model uses only information relating to the basic scoring dynamics within a sport, and knows nothing about individual teams or game strategies. In that light, its accuracy is impressive.

These results suggest several interesting directions for future work. For instance, further elucidating the connection between team skill and the observed scoring patterns would

[4] See `https://live.bovada.lv`. Only data on NBA games were available.
[5] See `http://www.sbrforum.com/betting-odds`

provide an important connection between within-game dynamics and team-specific characteristics. These, in turn, could be estimated from player-level characteristics to provide a coherent understanding of how individuals cooperate to produce a team and how teams compete to produce dynamics. Another missing piece of the dynamics puzzle is the role played by the environment and the control of space for creating scoring opportunities. Recent work on online games with heterogeneous environments suggests that these spatial factors can have large impact on scoring tempo and balance [81], but time series data on player positions on the field would further improve our understanding. Finally, our data omit many aspects of gameplay, including referee calls, timeouts, fouls, etc., which may provide for interesting strategic choices by teams, e.g., near the end of the game, as with clock management in football games. Progress on these and other questions would shed more light on the fundamental question of how much of gameplay may be attributed to skill versus luck.

Finally, our results demonstrate that common patterns and processes do indeed cut across seemingly distinct sports, and these patterns provide remarkably accurate descriptions of the events within these games and predictions of their outcomes. However, many questions remain unanswered, particularly as to what specific mechanisms generate the modest deviations from the basic patterns that we observe in each sport, and how exactly teams exerting such great efforts against each other can conspire to produce gameplay so reminiscent of simple stochastic processes. We look forward to future work that further investigates these questions, which we hope will continue to leverage the powerful tools and models of dynamical systems, statistical physics, and machine learning with increasingly detailed data on competition.

Chapter 7

Occupation networks: career path dynamics and income traps

7.1 Introduction

Career path dynamics and poverty traps are the product of complex interactions taking place between individual workers and their environment. Conventionally, these dynamics are studied macroscopically, using labor pools of millions of workers and economic variables, such as gross domestic product, representing the value of billions of goods and services [71]. While this low resolution analysis produces population level trends about an economy as a whole, it hides important information about how the independent, micro-level actions of individuals and changes to the environment produce these macro-level dynamics.

Network science continues to play an increasingly important role in understanding the complex dynamics of human behavior [35, 69, 84]. Its theoretical foundations provide a principled set of tools with which to perform scientific research [94]. However, these tools have largely been overlooked for answering fundamental questions about individual career path and economic dynamics, although see [56, 78].

Using a novel data set of nearly 1,000,000 publicly available resumes, 6,000,000 occupation descriptions, and 40,000 skills, we construct a rich network of occupations and study its structural properties. Our analysis indicates that nodes of high degree require social and administrative skills while low degree nodes require more specialized, technical skills. We also find that the number of common skills shared by a node and its neighbors decreases with degree, suggesting that a few core skills allow a worker to access many occupations. Finally,

we identify a region of the network, representing middle and upper-middle class incomes, that is vulnerable to income trap conditions, where the mean salary of a node's neighbors is lower than its own. We conclude with a brief discussion of how these results can be used to inform career path and economic policy decisions.

7.2 Data curation and network construction

We acquired 841,552 publicly available resumes from the website, `www.indeed.com`. Each resume was structured as an HTML document with specific tags encoding a worker's occupation title, dates worked and company name. We also acquired 6,069,636 occupation descriptions by crawling both `www.indeed.com` and `simplyhired.com`. Finally, we collected a set of 40,660 crowd sourced skills from LinkedIn's skills directory.

After de-aliasing the data, we constructed a network of occupations, where each vertex in the network is a unique occupation title observed at least twice in the set of resumes. Directed edges were placed between two vertices, a and b, if in a resume, a worker transitioned from occupation title a to occupation title b. This procedure produced a network containing 169,840 vertices inter-connected by 956,735 edges. Each vertex had an average in- and out-degree of 5.6.

Each vertex was annotated with an estimated average national salary, a set of skills, education level, and experience level. Salary annotations were obtained by searching a salary database provided by `www.indeed.com`. The remaining skill, education, and experience annotations were obtained by turning the occupation descriptions into sets of n-grams and then searching for phrases that matched those in the sets of skills, education levels, and years of experience.

The resumes contained in our data were collected during the months of April and May in the year 2010. These resumes were likely from workers actively seeking new positions, possibly as a result of being laid off due to the aftermath of the financial crisis in 2008. Additionally, the resumes represent a fraction of the population that had access to the

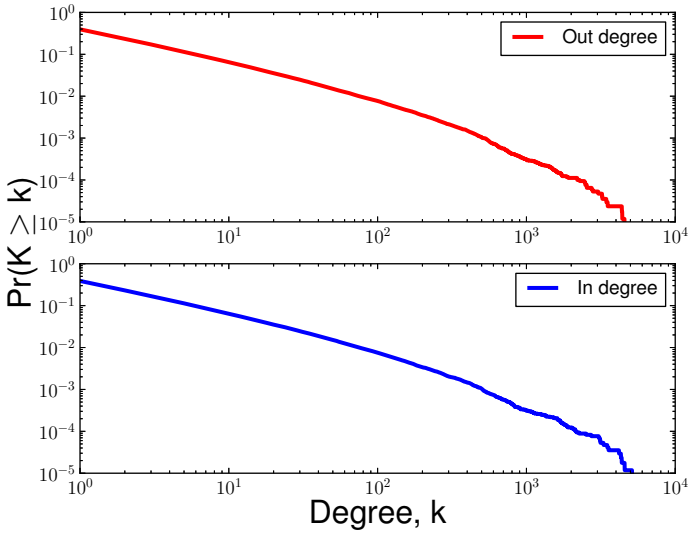

Figure 7.1: Complementary cumulative in- and out-degree distributions

Internet and knowledge of the resume service offered by Indeed. These idiosyncrasies almost certainly introduce a bias into our sample.

7.3 Career path dynamics

In this section, we analyze the structure of the occupation network in order to develop insights that may inform career path decisions. We begin by studying the network's degree distributions. Figure 7.1 plots the complementary cumulative degree in- and out-distributions. The heavy tailed structure in both plots indicates that a small fraction of vertices in the network have a surprisingly large number of occupation titles that lead to and from them.

We study the degree centralities of the network to determine if the high in-degree vertices are also the same vertices with high out-degrees. Tables 7.1 and 7.2 show the sets containing the 10 highest in-degree and out-degree vertices are nearly identical, differing by only a single occupation title. This similarity indicates that these vertices act as hubs in the network and that many seemingly distinct career trajectories cross paths at these titles.

occupation title	in-degree
owner	4759
administrative assistant	4559
manager	4401
customer service representative	4373
sales associate	3512
project manager	3440
office manager	3235
supervisor	3219
cashier	3035
assistant manager	2788

Table 7.1: Top 10 vertices with largest in-degree

occupation title	out-degree
administrative assistant	5174
customer service representative	5122
manager	4567
owner	4365
sales associate	4312
cashier	4185
assistant manager	3559
office manager	3491
supervisor	3295
receptionist	3140

Table 7.2: Top 10 vertices with largest out-degree

The occupation title with highest in-degree, owner, indicates that workers from a diverse set of occupation titles, transition into a business owner role at some point in their career. In addition, Table 7.2 indicates that owner is the vertex with the fourth largest out-degree. This pattern of high in- and out- degree suggests that many workers try to establish and run a small business, but as is well known, often times fail and transition back to working for an existing firm.

A qualitative comparison of these occupation titles suggests that positions with high degree require skills that are social in nature. That is, they require generic skills that many people develop through the processes of social conditioning and maturation. For example, seven skills shared by all top ten occupation titles are: customer service, procurement, conflict resolution, communication, sales, and business.

7.3.1 Skill assortativity

Next, we study how skill assortativity varies with degree. We measure skill assortativity by computing the Jaccard Similarity of skill sets between neighboring vertices [66]. Mathematically, Jaccard Similarity is defined as

$$J(\{A\},\{B\}) = \frac{|\{A\} \cap \{B\}|}{|\{A\} \cup \{B\}|} \tag{7.1}$$

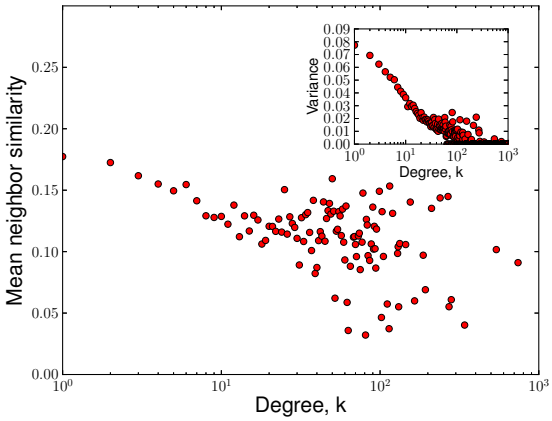

Figure 7.2: Mean neighbor skill similarity as a function of degree, bin size of 10. (inset) Variance of mean neighbor skill similarity.

where $\{A\}$ and $\{B\}$ are the skill sets of vertices A and B. A high similarity value indicates the vertices share many of the same skills while a low similarity value indicates the opposite. Figure 7.3 plots mean neighbor skill similarity as a function of degree. Vertices with low degree tend to share more skills with their neighbors than those with high degree. This pattern, along with titles contained in Table 7.1 and 7.2, suggests that low degree occupation titles are likely specialized roles that require higher levels of training or education, while high degree vertices are those that require less specialized skills.

7.3.2 Diversity

Next we study how degree varies with mean clustering coefficient. This relationship provides insight into an occupation title's level of diversity. Figure 7.2 plots mean local clustering as a function of degree for a random graph, generated by the configuration model with degree sequence equivalent to the undirected occupation network, and of the undirected occupation network. The figure indicates that low degree vertices in the occupation network have higher local clustering than in the random network. Vertices with high degree exhibit lower clustering than the random graph. The pattern of high local clustering at low degree

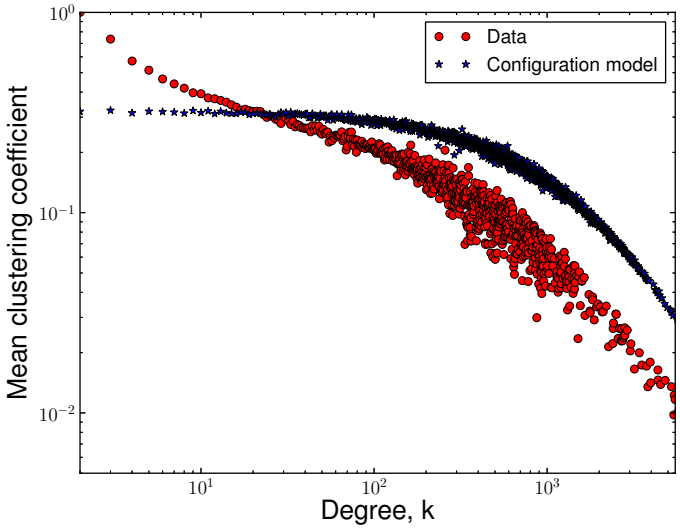

Figure 7.3: Mean local clustering as a function of degree for the undirected occupation network and a random network, generated with the configuration model and the undirected occupation network's degree sequence.

and low local clustering at high degree is likely due to the skill assortativity pattern, where occupation titles with low degree require high levels of training and occupation titles with high degree require less. Thus, high degree vertices fan out to a diverse set of occupations while low degree vertices tend to form interconnected groups of occupations which are more closely related in skill.

7.4 Income traps

Poverty traps are defined as self-reinforcing mechanisms which cause poverty to persist over time [8, 23]. Poverty trap mechanisms are classified into one of three primary groups, known as threshold effects, institutional effects, and neighborhood effects. Threshold models explain persistent poverty by identifying initial conditions, such as low life expectancy and capital, that alter the way a normal economy functions, such that the return on capital investments are small or negative. Institutional effects occur when governments obstruct the flow and allocation of capital such that only a small fraction of the population receives

enough to develop and grow. Finally, neighborhood effects are those which are caused by social contracts between family or other members of the community. Under these contracts, capital that is originally allocated to a single individual is re-allocated amongst a larger number of family or community members.

Traditionally, poverty traps are studied dynamically, using a deterministic difference equation of the form

$$x_{t+1} = F(x_t), \tag{7.2}$$

where $F(x)$ is a pre-defined growth model, such as the endogenous growth or Solow growth models [101, 110]. Most commonly, x_t is an aggregate measure, such as gross domestic product (GDP). Once the economy is initialized to x_0, its evolution can be studied by iteratively applying x_t as input to $F(x)$ and obtaining x_{t+1}. However, analyzing an economy at such a low resolution hides information about how the micro-actions of individuals produce the macro-patterns of the economy.

To study similar patterns in individual salary dynamics, which we refer to as income traps, using the directed occupation network, we analyze changes between a vertex's salary and adjacent neighbor salaries. Figure 7.4(Left) plots the salaries x, rounded to the nearest \$10,000 and the mean salary of all adjacent occupations, $x + 1$. The error bars surrounding each point indicate the standard deviation. From this figure, it is clear that three ranges exist within the occupation network, which we will refer to as low, middle, and high incomes.

Low incomes - Workers with occupation titles earning less than \$50,000 at position x typically earn \$50,000 when they move to their next position. All subsequent occupation changes earn on average \$50,000. This indicates that workers in the low range should never expect to earn more than \$50,000 over the course of their careers.

Middle incomes - Workers with occupation titles earning between \$80,000 and \$180,000 form an income trap. On average, workers earning more than \$80,000 and less than \$180,000 take income losses after each occupation title change. The income trap is lower bounded by

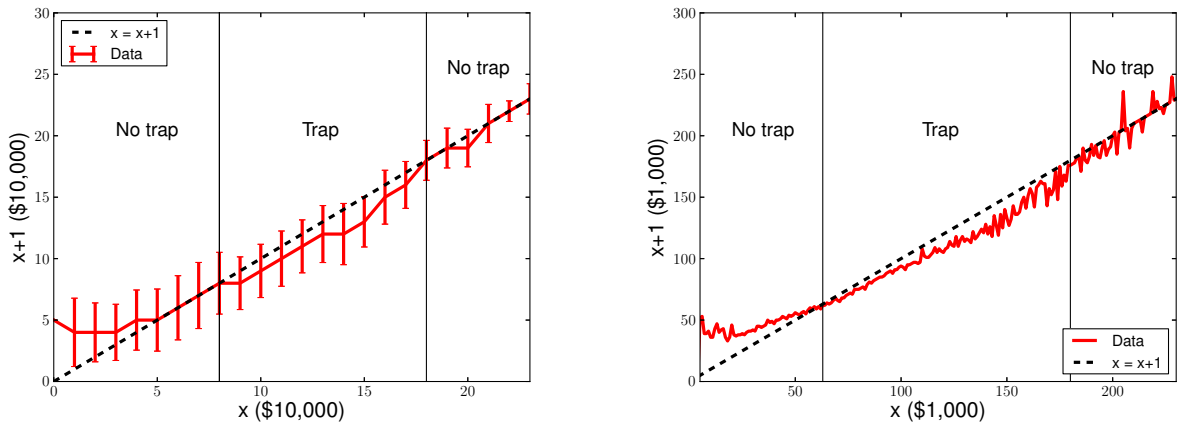

Figure 7.4: (Left) Income trap map with incomes rounded to the nearest $10,000 (Right) Income trap map, with incomes rounded to the nearest $1,000.

$80,000, indicating the workers in this region will never earn less than $80,000, even after changing occupation titles many times.

High incomes - Workers earning more than $180,000 per year in salary tend to earn roughly the same income after each occupation title change, although workers earning $200,000 tend to take a $10,000 decrease in salary, upon occupation title change.

The structure of the income map indicates that the workers earning less than $80,000 will earn between $50,000 and $80,000 per year over the course of their careers. Similarly, workers that begin earning more than $80,000 and less than $180,000 will earn no less than $80,000 annually over the course of their careers.

To test the robustness of the data, we also study the map produced by rounding salaries to the nearest $1,000. As shown in Figure 7.4(Right), we find the same pattern, although the the lower bound on the middle income trap is lower, beginning at $63,000 instead of $80,000.

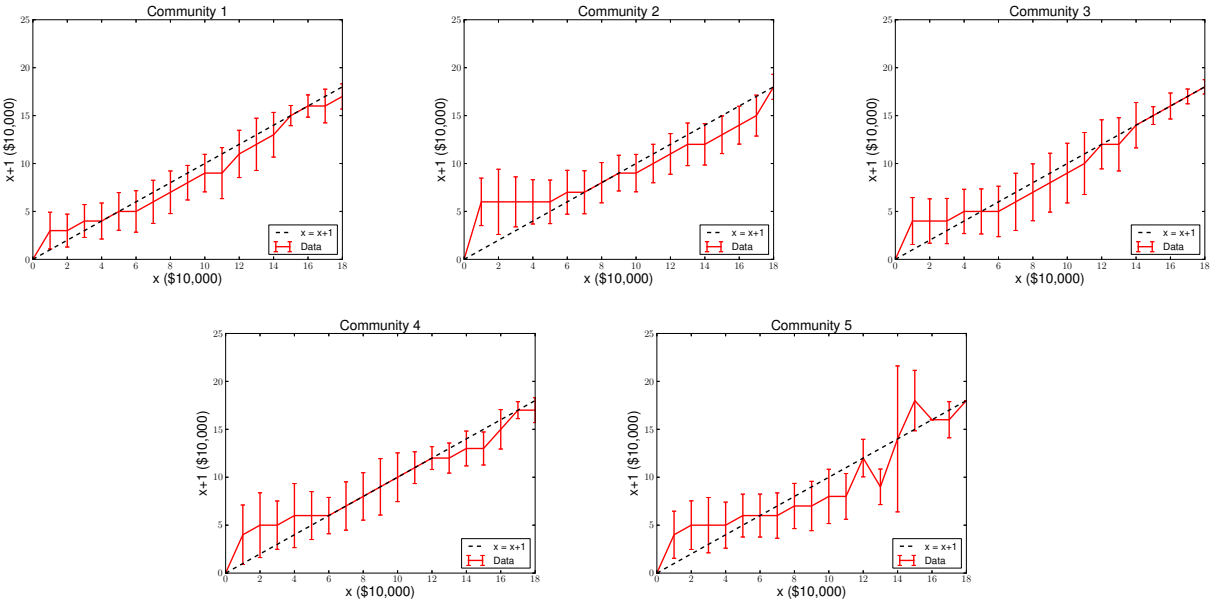

Figure 7.5: Income trap maps with incomes rounded to the nearest $10,000 for the 5 largest communities in the network.

7.5 Communities

Previously, we studied income dynamics at the population level. Here we study how patterns in income traps vary by community in the occupation network. We identify communities using the most popular unsupervised method for extracting communities from a network, modularity maximization, where modularity is measured as the fraction of edges that connect vertices in a module less the expected number of edges that would connect the vertices if the network were purely random [32].

The five largest communities, which account for roughly 85% of vertices in the network, exhibit income patterns similar to those identified in the global income map (see Fig. 7.5). Across all communities, we find occupations with salaries below $40,000 are typically followed by an occupation earning between $40,000 and $50,000. As observed in the population level map, a threshold between $40,000 and $70,000, depending on community, acts as an upper bound on maximum salary.

Income patterns falling in the range between $50,000 and $150,000 vary by community.

In communities 1 and 3, an income trap exists between incomes of \$60,000 \$150,000 and \$50,000 and \$120,000 respectively. In community 2, a similar trap exists, but begins at \$90,000. Surprisingly, community 4 does not exhibit an income trap pattern in this income range. Community 5 exhibits the most severe income trap of all communities, with the trap pattern beginning at roughly \$60,000 and persisting through nearly all higher incomes.

High incomes, those greater than roughly \$150,000, in communities 1 and 3 are robust to occupation title changes, meaning that workers earning an income greater than or equal to \$150,000 will tend earn at least this much in all future occupation title changes. However, in the remaining communities (2,4,5) these incomes exhibit the same trap pattern as found in the middle income range.

In nearly all communities, career paths containing occupation titles with earnings between \$50,000 and \$150,000 suffer from income traps. That is, in this salary range, over the course of a career, a worker's salary is likely to degrade with each job change. This pattern supports the popular belief that the middle class is and will continue to decrease in size.

7.6 Long term income distribution

To study the long term effects of this dynamical system on the distribution of worker incomes, we perform a non-parametric simulation of workers and their career paths. We simulate 1,000,000 workers, each of who change occupation titles 10 times, representing a common total number of job changes per career in the industrial work force. For each simulated worker, we choose a starting vertex in the network uniformly at random and then choose the next occupation title from the vertex's set of outgoing edges, again uniformly at random. The salary of the simulated worker is recorded after repeating this process 10 times.

The income maps shown in Figures 7.4 and 7.5 suggest that most workers will end up earning roughly \$50,000. Indeed, Figure 7.6 plots the simulated distribution of incomes and indicates that an overwhelming majority of the population of workers earn slightly less

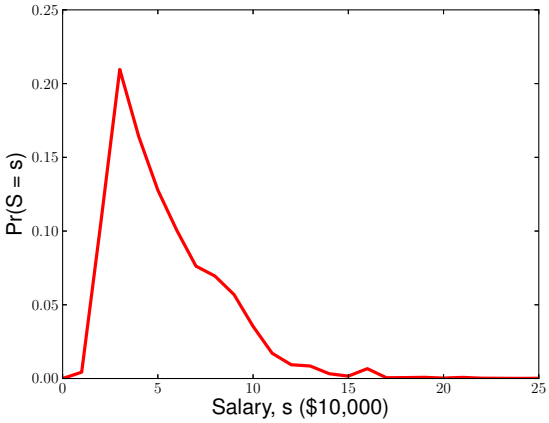

Figure 7.6: Distribution of incomes after simulating 1,000,000 workers, whose initial incomes are chosen uniformly at random, changing occupation titles 10 times over the course of their careers.

than $50,000 per year while a minuscule fraction earn more than $180,000, a pattern that supports the theory that the middle class of the United States is indeed shrinking and the lower class is expanding.

7.7 Discussion

Applying a network centric approach to the study of career path dynamics and income traps enables one to understand how the micro-actions of individuals produce commonly studied macro-level patterns such as changes in GDP. By understanding network structure and dynamics, individuals and policy makers can make better informed decisions.

From a career path point of view, we identified several important structural features that provide individuals with strategic information about how to plan their career paths. Specifically, occupation titles with high in- or out-degrees possess skill sets that are socially oriented and emphasize customer service, business management and sales.

We also found that the skill sets of adjacent vertices tend to be less alike with increasing degree, a pattern suggesting that low degree occupation titles are likely specialized roles that require higher levels of training or education, while high degree vertices are those that require

less specialized skills.

Low degree vertices also tend to have higher local clustering coefficients, further suggesting that low degree vertices connect to and form tightly knit groups that share skills. High degree vertices tend to connect to a diverse set of occupation titles with little similarity, thus acting as a hub in the network and connecting to a wide variety of loosely related occupation titles.

At the population and community levels, income traps pervade the occupation network, primarily in the middle class range of income levels. When simulating the long term distribution of incomes generated by this pattern, we find that an overwhelming majority of workers earn a salary of roughly $40,000 when they stop working, even if they earned a higher salary earlier in their career. This pattern supports the current belief that the middle class is indeed shrinking.

Our work has only explored a subset of the occupation network's structural and dynamical properties. Further developing the relationship between vertex degree, skill set, education level and earnings would allow for a deeper understanding of how these features are related to one another and how these patterns can be used to make career path decisions that maximize various objectives, such as income, and economic policy decisions to mitigate the observed middle class income traps. We look forward to future studies exploring these and other dynamical and structural aspects of occupation networks.

Chapter 8

Conclusion

In this thesis, we have studied three distinct competitive social systems, a massive online game, a set of professional sports and the labor market. In each system, we discovered how statistical patterns in group and individual level behavior can be used to model and in many cases, predict dynamics and outcomes.

In Chapters 3 and 4, we identified a set of temporal and pro-social features that can be used to make highly accurate inferences about whether two individuals are friends within the online game Halo: *Reach*. Using this predictive model, we inferred the social network embedded in the game and studied how it evolved over time. We found that some of its dynamical properties contradict many popular beliefs and theories about online social networks. In particular, we found that the social network within *Reach* does not become more dense over time, or with increasing size.

In Chapter 5, we developed a generative model of competition that accurately captures the timing and scoring dynamics of competition in Halo: *Reach*. Moreover, we developed a model that connected variations in these dynamics with variations in the environmental structure of a competition. These models contribute towards the development of a quantitative approach to designing online games and capturing how variation in a competition's environment influences its dynamics.

In Chapter 6, we extended our model of competition, developed in Chapter 5, and applied it to the timing and scoring properties of three different professional sports: football,

hockey and basketball. We discovered that the timing of scoring events in these games is well represented by a homogenous Poisson process. We also found that patterns in scoring can be modeled as a Bernoulli process whose bias parameter varies with the difference in score. Using these properties, we constructed a Markov chain on the difference in score and were able to accurately predict outcomes in an on-line fashion as a game evolved.

Chapter 7, we constructed and analyzed a network of occupations. Structural properties of the network such as degree centrality and local clustering, combined with additional information such as an occupation's skill set, indicate that occupations with high in- and out-degrees tend to act as a hubs, inter-connecting many disparate career paths that rely on a small set of socially oriented skills, such as customer service and sales. We also studied the income dynamics of workers on the occupation network and found that workers earning anywhere between roughly $50,000 and $180,000 suffer from income trap dynamics, where their next occupation pays less than their current one.

While this thesis has answered many questions about competitive dynamics and behavioral patterns in competitive social systems, it has also raised new ones. In *Reach*, comparing and contrasting the dynamics of the inferred social network the dynamics of the system's interaction network would deepen and broaden our understanding of how friendship dynamics and pure interaction dynamics differ from one another.

In professional sports, building a mathematical model that connects the team-level parameters of our model of competition to player-level performance parameters would complete our top-down approach to understanding competitive dynamics in sports.

And finally, analyzing the community structure of the occupation in finer detail, specifically identifying vertices that act as bridges, can shed light on which occupations are trans-industrial and thus act as a sort of glue, bringing together what would normally be islands distinct occupations.

We look forward to future studies on competitive dynamics in complex social systems.

Bibliography

[1] Lada A Adamic and Eytan Adar. Friends and neighbors on the web. Social Networks, 25(3):211–230, 2003.

[2] Muhammad Aurangzeb Ahmad, Brian Keegan, Jaideep Srivastava, Dmitri Williams, and Noshir Contractor. Mining for gold farmers: Automatic detection of deviant players in MMOGs. In Conf. on Comp. Science and Engineering, pages 340–345. IEEE, 2009.

[3] Yong-Yeol Ahn, Seungyeop Han, Haewoon Kwak, Sue Moon, and Hawoong Jeong. Analysis of topological characteristics of huge online social networking services. In Proc. World Wide Web, pages 835–844. ACM, 2007.

[4] Jim Albert, Jay Bennett, and James J Cochran. Anthology of statistics in sports, volume 16. Society for Industrial Mathematics, 2005.

[5] Jeremy Arkes and Jose Martinez. Finally, evidence for a momentum effect in the nba. Journal of Quantitative Analysis in Sports, 7(3):1–14, 2011.

[6] Entertainment Software Association. Essential facts about the compute and video game industry, 2011.

[7] Peter Ayton and Ilan Fischer. The hot hand fallacy and the gamblers fallacy: Two faces of subjective randomness? Memory & Cognition, 32(8):1369–1378, 2004.

[8] Costas Azariadis and John Stachurski. Poverty traps. Handbook of economic growth, 1:295–384, 2005.

[9] Lars Backstrom, Paolo Boldi, Marco Rosa, Johan Ugander, and Sebastiano Vigna. Four degrees of separation. In Proc. Web Science, pages 33–42. ACM, 2012.

[10] Eytan Bakshy, Matthew P Simmons, David Huffaker, Chun-Yuen Teng, and Lada Adamic. The social dynamics of economic activity in a virtual world. Proc. Weblogs and Social Media, 1001:48103, 2010.

[11] Timothy T Baldwin, Michael D Bedell, and Jonathan L Johnson. The social fabric of a team-based M.B.A. program: Network effects on student satisfaction and performance. Academy of Management Journal, 40:1369–1397, 1997.

[12] Prasad Balkundi and David A Harrison. Ties, leaders, and time in teams: Strong inference about network structure's effects on team viability and performance. Academy of Management Journal, 49:49–68, 2006.

[13] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. Science, 286(5439):509–512, 1999.

[14] Jay Barney. Firm resources and sustained competitive advantage. Journal of Management, 17:99–120, 1991.

[15] Eli Ben-Naim, NW Hengartner, Sidney Redner, and Federico Vazquez. Randomness in competitions. Journal of Statistical Physics, pages 1–17, 2012.

[16] Eli Ben-Naim, Federico Vazquez, and Sidney Redner. What is the most competitive sport? Journal of the Korean Physical Society, 50:124, 2007.

[17] Jonah Berger and Devin Pope. Can losing lead to winning? Management Science, 57(5):817–827, 2011.

[18] Christopher M Bishop. Pattern Recognition and Machine Learning, volume 4. Springer New York, 2006.

[19] Jeremy Blackburn, Ramanuja Simha, Nicolas Kourtellis, Xiang Zuo, Matei Ripeanu, John Skvoretz, and Adriana Iamnitchi. Branded with a scarlet C: cheaters in a gaming social network. In Proc. World Wide Web, pages 81–90. ACM, 2012.

[20] Mary L Boas. Mathematical Methods in the Physical Sciences. John Wiley & Sons, Inc., Hoboken, NJ, 3 edition, 2006.

[21] Douglas H. Boucher. The idea of mutualism, past and future. In Douglas H. Boucher, editor, The Biology of Mutualism: Ecology and Evolution, pages 1–27. Oxford University Press, 1985.

[22] Samuel Bowles. Microeconomics: Behavior, Institutions, and Evolution. Princeton University Press, 2006.

[23] Samuel Bowles, Steven N Durlauf, and Karla Hoff. Poverty traps. Princeton University Press, 2006.

[24] George EP Box, Gwilym M Jenkins, and Gregory C Reinsel. Time series analysis: forecasting and control, volume 734. Wiley, 2011.

[25] Andrew P Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition, 30(7):1145–1159, 1997.

[26] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. Biometrika, 39(3/4):324–345, 1952.

[27] Zdzisław Brzeźniak and Tomasz Jerzy Zastawniak. Basic Stochastic Processes. Springer, Berlin, 2000.

[28] Samuel E Buttrey, Alan R Washburn, and Wilson L Price. Estimating NHL scoring rates. Journal of Quantitative Analysis in Sports, 7(3), 2011.

[29] Edward Castronova, Dmitri Williams, Cuihua Shen, Rabindra Ratan, Li Xiong, Yun Huang, and Brian Keegan. As real as real? macroeconomic behavior in a large-scale virtual world. New Media & Society, 11(5):685, 2009.

[30] Aaron Clauset and Nathan Eagle. Persistence and periodicity in a dynamic proximity network. In DIMACS Workshop on Computational Methods for Dynamic Interaction Networks, pages 1–5, 2007.

[31] Aaron Clauset, Cris Moore, and Mark EJ Newman. Hierarchical structure and the prediction of missing links in networks. Nature, 453(7191):98–101, 2008.

[32] Aaron Clauset, Mark EJ Newman, and Cris Moore. Finding community structure in very large networks. Physical review E, 70(6):066111, 2004.

[33] Aaron Clauset, Cosma R Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. SIAM Review, 51:661–703, 2009.

[34] Charles Horton Cooley. Social organization: A study of the larger mind. Transaction Publishers, 1983.

[35] David J Crandall, Lars Backstrom, Dan Cosley, Siddharth Suri, Daniel Huttenlocher, and Jon Kleinberg. Inferring social ties from geographic coincidences. Proc. National Academy of Sciences, 107(52):22436–22441, 2010.

[36] Justin Cranshaw, Eran Toch, Jason Hong, Aniket Kittur, and Norman Sadeh. Bridging the gap between physical location and online social networks. In Proc. Ubiquitous Computing. ACM, September 2010.

[37] Allison Davis, Burleigh B Gardner, and Mary R Gardner. Deep south. University of Chicago Press Chicago, 1941.

[38] Munmun De Choudhury, Winter A Mason, Jake M Hofman, and Duncan J Watts. Inferring relevant social networks from interpersonal communication. In Proc. World Wide Web, pages 301–310. ACM, 2010.

[39] Y. de Saá Guerra, J.M. Martín González, S. Sarmiento Montesdeoca, D. Rodríguez Ruiz, N. Arjonilla López, and J.M. García Manso. Basketball scoring in nba games: an example of complexity. Journal of Systems Science and Complexity, 26(1):94–103, 2013.

[40] Derek de Solla Price. A general theory of bibliometric and other cumulative advantage processes. Journal of the American Society for Information Science, 27(5):292–306, 1976.

[41] Derek J de Solla Price. Networks of scientific papers. Science, 149:510–515, 1965.

[42] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: a flexible data processing tool. Communications of the ACM, 53(1):72–77, 2010.

[43] Jerker Denrell. Random walks and sustained competitive advantage. Management Science, 50:922–934, 2004.

[44] Jerker Denrell and Chengwei Liu. Top performers are not the most impressive when extreme performance indicates unreliability. Proc. National Academy Sciences, 109:9331–9336, 2012.

[45] Nathan Eagle, Alex Sandy Pentland, and David Lazer. Inferring friendship network structure by using mobile phone data. Proc. National Academy of Sciences, 106(36):15274–15278, 2009.

[46] Bradley Efron and Robert Tibshirani. An Introduction to the Bootstrap. Chapman & Hall, New York, NY, 1993.

[47] Entertainment Software Association. Essential Facts about the Computer and Video Game Industry. http://bit.ly/kLHJ2Q, (access date February, 2012), 2011.

[48] Paul Erdos and Alfred Renyi. On random graphs I. Publ. Math. Debrecen, 6:290–297, 1959.

[49] Phil Everson and Paul S Goldsmith-Pinkham. Composite poisson models for goal scoring. Journal of Quantitative Analysis in Sports, 4(2), 2008.

[50] Alan Gabel and Sidney Redner. Random walk picture of basketball scoring. Journal of Quantitative Analysis in Sports, 8:Manuscript 1416, 2012.

[51] Tobias Galla and J Doyne Farmer. Complex dynamics in learning complicated games. Proc. National Academy Sciences, 110:1232–1236, 2013.

[52] David Garcia, Pavlin Mavrodiev, and Frank Schweitzer. Social resilience in online communities: The autopsy of friendster. arXiv preprint arXiv:1302.6109, 2013.

[53] Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung. The google file system. In ACM SIGOPS Operating Systems Review, volume 37, pages 29–43. ACM, 2003.

[54] Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. Proc. National Academy of Sciences, 99(12):7821–7826, 2002.

[55] Matt Goldman and Justin M Rao. Effort vs. concentration: The asymmetric impact of pressure on NBA performance. In Proc. MIT Sloan Sports Analytics Conference, pages 1–10, 2012.

[56] Omar A Guerrero and Robert L Axtell. Employment growth through labor flow networks. PloS one, 8(5):e60808, 2013.

[57] Roger Guimerà, Brian Uzzi, Jarrett Spiro, and Luís A. Nunes Amaral. Team assembly mechanisms determine collaboration network structure and team performance. Science, 308(5722):697–702, 2005.

[58] Nobuyuki Hanaki, Alexander Peterhansl, Peter S Dodds, and Duncan J Watts. Cooperation in evolving social networks. Management Science, 53(7):1036–1050, 2007.

[59] James A Hanley. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 143:29–36, 1982.

[60] Ralf Herbrich, Tom Minka, and Thore Graepel. Trueskill^{TM}: A bayesian skill rating system. Advances in Neural Information Processing Systems, 19:569, 2007.

[61] Andreas Heuer and Oliver Rubner. How does the past of a soccer match influence its future? Preprint, http://arxiv.org/abs/1207.4471, 2012.

[62] Jeff Huang, Thomas Zimmermann, Nachiappan Nagapan, Charles Harrison, and Bruce C Phillips. Mastering the art of war: how patterns of gameplay influence skill in halo. In Proc. Conf. on Human Factors in Computing Systems, pages 695–704. ACM, 2013.

[63] Yun Huang, Cuihua Shen, Dmitri Williams, and Noshir Contractor. Virtually there: Exploring proximity and homophily in a virtual world. In Conf. on Comp. Science and Engineering, volume 4, pages 354–359. IEEE, 2009.

[64] Yun Huang, Mengxiao Zhu, Jing Wang, Nishith Pathak, Cuihua Shen, Brian Keegan, Dmitri Williams, and Noshir Contractor. The formation of task-oriented groups: Exploring combat activities in online games. In Conf. on Comp. Science and Engineering, pages 122–127. IEEE, 2009.

[65] David Huffaker, Jing Wang, Jeffrey Treem, Muhammad Aurangzeb Ahmad, Lindsay Fullerton, Dmitri Williams, Marshall Scott Poole, and Noshir Contractor. The social behaviors of experts in massive multiplayer online role-playing games. In Conf. on Comp. Science and Engineering, volume 4, pages 326–331. IEEE, 2009.

[66] Paul Jaccard. The distribution of the flora in the alpine zone. 1. New phytologist, 11(2):37–50, 1912.

[67] Carole Sève Jérôme Bourbousson and Tim McGarry. Space-time coordination dynamics in basketball: Part 2. The interaction between the two teams. Journal of Sports Sciences, 28(3):349–358, 2012.

[68] Jason J Jones, Jaime E Settle, Robert M Bond, Christopher J Fariss, Cameron Marlow, and James H Fowler. Inferring tie strength from online directed behavior. Plos one, 8(1):e52168, 2013.

[69] Brian Keegan, Muhammad Aurangzeb Ahmed, Dmitri Williams, Jaideep Srivastava, and Noshir Contractor. Dark gold: Statistical properties of clandestine networks in massively multiplayer online games. In Conf. on Social Computing, pages 201–208. IEEE, 2010.

[70] Eliza Kern. Facebook is collecting your data 500 terabytes a day. `http://gigaom.com/2012/08/22/facebook-is-collecting-your-data-500-terabytes-a-day/`, 2012. [Online; accessed 15-April-2013].

[71] Paul Krugman, Robin Wells, and Martha Olney. Essentials of economics. Macmillan, 2006.

[72] Ravi Kumar, Jasmine Novak, and Andrew Tomkins. Structure and evolution of online social networks. In Proc. Knowledge discovery and data mining, pages 611–617, New York, NY, USA, 2006. ACM.

[73] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In Proc. World Wide Web, pages 591–600. ACM, 2010.

[74] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. Predicting positive and negative links in online social networks. In Proc. World Wide Web, pages 641–650. ACM, 2010.

[75] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In Proc. Knowledge discovery in data mining, pages 177–187. ACM, 2005.

[76] Xin Li, Lei Guo, and Yihong Eric Zhao. Tag-based social interest discovery. In Proc. World Wide Web, pages 675–684. ACM, 2008.

[77] David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. Journal of the American Society for Info. Science and Technology, 58(7):1019–1031, 2007.

[78] Eliza-Olivia Lungu, Ana-Maria Zamfir, and Cristina Mocanu. Occupational mobility networks of female and male higher education graduates. Theoretical and Applied Economics, 18(8 (585)):37–46, 2013.

[79] Winter Mason and Aaron Clauset. Friends ftw! friendship and competition in halo: Reach. In Proc. computer supported cooperative work, pages 375–386. ACM, 2013.

[80] Winter Mason and Aaron Clauset. Friends FTW! Friendship and competition in *Halo: Reach*. In Proc. Computer Supported Coop. Work and Social Comp., 2013.

[81] Sears Merritt and Aaron Clauset. Environmental structure and competitive scoring advantages in team competitions. Scientific Reports, 3(3067), 2013.

[82] Sears Merritt and Aaron Clauset. Scoring dynamics across professional team sports: tempo, balance and predictability. arXiv preprint arXiv:1310.4461, 2013.

[83] Sears Merritt and Aaron Clauset. Social network dynamics in a massive online game: Network turnover, non-densification, and team engagement in halo reach. In Proc. Workshop on Mining and Learning from Graphs, 2013.

[84] Sears Merritt, Abigail Z Jacobs, Winter Mason, and Aaron Clauset. Detecting friendship within dynamic online interaction networks. In Proc. Weblogs and Social Media, 2013.

[85] David R Michael and Sandra L Chen. Serious Games: Games That Educate, Train, and Inform. Muska and Lipman, 2005.

[86] Stanley Milgram. The small world problem. Psychology today, 2(1):60–67, 1967.

[87] Alan Mislove, Massimiliano Marcon, Krishna P Gummadi, Peter Druschel, and Bobby Bhattacharjee. Measurement and analysis of online social networks. In Proc. Internet measurement, pages 29–42. ACM, 2007.

[88] Roger B Myerson. Game Theory: Analysis of Conflict. Harvard University Press, Cambridge MA, 1997.

[89] National Basketball Association. Official rules of the national basketball association. 2013.

[90] National Collegiate Athletic Association. Football, rules and interpretations. 2013.

[91] National Football Association. Official rules of the national football association. 2013.

[92] National Hockey League. Official rules of the national hockey league. 2013.

[93] Tal Neiman and Yonatan Loewenstein. Reinforcement learning in professional basketball players. Nature communications, 2:569, 2011.

[94] Mark EJ Newman. The structure and function of complex networks. SIAM review, 45(2):167–256, 2003.

[95] Mark EJ Newman. Networks: An Introduction. Oxford University Press, 2010.

[96] Mark EJ Newman, Steven H Strogatz, and Duncan J Watts. Random graphs with arbitrary degree distributions and their applications. Physical Review E, 64(2):026118, 2001.

[97] Noam Nisan. Algorithmic game theory. Cambridge University Press, 2007.

[98] Ignacio Palacios-Huerta. Professionals play minimax. The Review of Economic Studies, 70(2):395–415, 2003.

[99] David Reed and Mike Hughes. An exploration of team sport as a dynamical system. Journal of Performance Analysis in Sport, 6(2):114–125, 2006.

[100] David Romer. Do firms maximize? Evidence from professional football. Journal of Political Economy, 114(2):340–365, 2006.

[101] Paul M Romer. The origins of endogenous growth. The Journal of Economic Perspectives, 8(1):3–22, 1994.

[102] Martin Ruef, Howard E Aldrich, and Nancy M Carter. The structure of founding teams: Homophily, strong ties, and isolation among U.S. entrepreneurs. American Sociological Review, 68:195–222, 2003.

[103] Matthew J Salganik, Peter Sheridan Dodds, and Duncan J Watts. Experimental study of inequality and unpredictability in an artificial cultural market. Science, 311:854–856, 2006.

[104] Purnamrita Sarkar, Deepayan Chakrabarti, and Michael Jordan. Nonparametric link prediction in dynamic networks. In Proc. Machine Learning, 2012.

[105] Kyong Jin Shim, Samarth Damania, Colin DeLong, and Jaideep Srivastava. Player and team performance in Everquest II and Halo 3. IEEE Potentials, 30(5), 2011.

[106] Kyong Jin Shim, Kuo-Wei Hsu, Samarth Damania, Colin DeLong, and Jaideep Srivastava. An exploratory study of player and team performance in multiplayer first-person-shooter games. In Conf. on Social Computing, pages 617–620. IEEE, 2011.

[107] Kyong Jin Shim, Richa Sharan, and Jaideep Srivastava. Player performance prediction in massively multiplayer online role-playing games (MMORPGs). Advances in Knowledge Discovery and Data Mining, pages 71–80, 2010.

[108] Kyong Jin Shim and Jaideep Srivastava. Team performance prediction in massively multiplayer online role-playing games (MMORPGs). In Conf. Social Computing, pages 128–136. IEEE, 2010.

[109] Clément Sire and Sidney Redner. Understanding baseball team standings and streaks. The European Physical Journal B, 67(3):473–481, 2009.

[110] Robert M Solow. A contribution to the theory of economic growth. The quarterly journal of economics, 70(1):65–94, 1956.

[111] Siddharth Suri and Sergei Vassilvitskii. Counting triangles and the curse of the last reducer. In Proc. World Wide Web, pages 607–614. ACM, 2011.

[112] Michael Szell, Renaud Lambiotte, and Stefan Thurner. Multirelational organization of large-scale social networks in an online world. Proc. National Academy of Sciences, 107(31):13636–13641, 2010.

[113] Andrew C Thomas. Inter-arrival times of goals in ice hockey. Journal of Quantitative Analysis in Sports, 3(3), 2007.

[114] Peter Thompson. Learning by doing. In Bronwyn Hall and Nathan Rosenberg, editors, Handbook of Economics of Technical Change, pages 429–476. Elsevier/North-Holland, 2010.

[115] Johan Ugander, Brian Karrer, Lars Backstrom, and Cameron Marlow. The anatomy of the facebook social graph. arXiv preprint arXiv:1111.4503, 2011.

[116] Roger C Vergin. Winning streaks in sports and the misperception of momentum. Journal of Sport Behavior, 23, 2000.

[117] Mark Walker and John Wooders. Minimax play at wimbledon. The American Economic Review, 91(5):1521–1538, 2001.

[118] Larry Wasserman. All of statistics: a concise course in statistical inference. Springer Verlag, 2004.

[119] Duncan J Watts and Steven H Strogatz. Collective dynamics of small-world networks. Nature, 393(6684):440–442, 1998.

[120] Tom White. Hadoop: the definitive guide. O'Reilly, 2012.

[121] Shaomei Wu, Jake M Hofman, Winter A Mason, and Duncan J Watts. Who says what to whom on twitter. In Proc. World Wide Web, pages 705–714. ACM, 2011.

[122] Martin Zinkevich, Markus Weimer, Lihong Li, and Alex J Smola. Parallelized stochastic gradient descent. In Advances in Neural Information Processing Systems, pages 2595–2603, 2010.

Appendix A

Environmental structure and competitive scoring advantages in team competitions

Supporting Information

A.1 Detailed Description of Data

Halo: Reach is a popular online game played by nearly 20 million individuals, and was the 3rd most popular US video game of 2010 [47]. It was publicly released by Bungie Inc., a former subdivision of Microsoft Game Studios, on 14 September 2010, and since then, players have generated more than 1 billion competitions. *Reach* is an example of the kind of virtual combat simulation known as a "first-person shooter" or FPS. Within the *Reach* system, players choose from among roughly seven primary game types and numerous subtypes, which are played on more than 33 terrain maps with 74 weapons (the precise number of maps and weapons has varied over time, as the publisher has periodically revised the online content through downloadable updates).

Instances of the game can be played alone, with or against other players via the Xbox Live online system. Participation in this system requires an account, which is distinguished by unique and publicly known "gamertag" or online pseudonym, chosen by the player. In the *Reach* system, both individual game and player summaries were made publicly available through the Halo Reach Stats API. Through this digital interface, we collected detailed data on the first 53 million competition instances (roughly 1TB of data).

Within our sample, there are three basic game types: **campaign games**, a sequence

of story-driven, player-versus-environment (PvE) maps that many players complete first; **firefight games** (also PvE), in which a team of human-controlled players battle successive waves of computer-controlled enemies; and **competitive games**, a player-versus-player (PvP) game type, in which teams of the equal size (2, 4, 6 or 8 players) compete to either be the first to reach some fixed number of points or have the largest score after a fixed length of time. (The precise number of players per team, number of points required to win and length of a game depends on the game subtype.) Here, we focus on the most common type of competitive game, with teams of 4 players, a time limit of 600 seconds and a score limit of 50 points.

Among other information, each competition instance game file includes the sequence of scoring events at the per-second resolution and a list of players by team. Scoring events are annotated with the gamertag of the player generating the event, the number of points scored and the player giving up the points (if applicable).

Unlike professional sports, team composition and player resources in **Reach** competitions are not persistent across instances. The only attribute that persists is individual player skill, and thus each new instance is a kind of a "blank slate." To join a new instance, individual players or small groups (often friends [79]) first enter a general pool of available competitors. A Bayesian "matchmaking" algorithm, which seeks to build teams of equal skill [60], then fills teams in the new instance by drawing from this pool. This process substantially randomizes the pairing of individuals within teams and the pairing of teams across instances. Because of the matchmaking algorithm and the large size of the pools, a pair of non-friend players are highly unlikely to be paired again in a new instance; friends may elect to be matched as a unit by forming a "party," a special grouping that the matchmaking algorithm recognizes.

The non-persistence and the randomization are features absent from most studies of team performance or competition [102, 11, 12], and serve to mitigate the confounding effects of persistent teams and resources present in most competitive systems, e.g., professional

sports. For our purposes, these features make *Reach* competitions a unique source of data for studying behavioral dynamics within competitions and how structural factors shape this behavior.

In competitive games, players move their avatars through the game map simultaneously, in real-time, navigating complex terrain, acquiring avatar modifications and encountering opponents. Teammates may interact through a private voice channel, or through visual signals. Points are scored by dealing sufficient damage to eliminate an opposing avatar and for each such success, a team gains a single point. Eliminated players must then wait several seconds before their avatar is placed back into the game at one of several specified "spawn" locations, equipped with "default" avatar resources that depend on the competition type being played.

For our analysis, we exclude all PvE games and all PvP games containing corrupt scoring event data. (Our analysis suggests no specific pattern to the corruption.) In our primary analyses, we further restricted our sample to PvP competitions (i) between two teams of 4 players and (ii) where no player exited the game early. This latter criterion was relaxed to calculate the relationship between dropouts and β (see Section A.8).

A.2 Generative Model for Scoring Event Timing and Balance

The timing and balance (which team receives the point) of scoring events within a competition are modeled by a conditionally independent Markov process, where an incremental change to a team's score s_r is given by

$$\Pr(\Delta s_r(t) > 0) = \Pr(\Delta s_r > 0 \,|\, \theta, \text{event}) \Pr(\text{event at } t \,|\, \theta) \ ,$$

where θ parameterizes the impact of non-ideal competitive features. That is, the probability that team r's score increases at some time t is the probability that a scoring event occurred at time t and that the resulting point was awarded to r. Furthermore, team labels r and b are arbitrary, and we choose r as our reference team below.

The generation of scoring events is given by a non-stationary Poisson process, in which the probability that a scoring event occurs at time t varies linearly with time:

$$\Pr(\text{event at } t \,|\, \lambda_0, \alpha\,) = \lambda_0 + \alpha\, t \ , \tag{A.1}$$

where λ_0 is the event background rate and α is the acceleration. When $\alpha = 0$, we recover the stationary Poisson process expected for ideal competitions.

In a real competition, we observe $n \leq T$ scoring events, for a competition lasting T units of time. Let $\{t_i\}$ denote the observed times of these events, and $\{u_j\}$ the times at which no event was observed. The model parameters λ_0 and α are then jointly estimated by directly maximizing the generative model's log-likelihood function:

$$\ln \mathcal{L} = \sum_{i=1}^{n} \ln(\lambda_0 + \alpha\, t_i) + \sum_{j=1}^{T-n} \ln(1 - \lambda_0 - \alpha\, u_j) \ . \tag{A.2}$$

To limit the biasing effect of the highly non-stationary behavior found in the early- and end-phases of competitions (see main text), we restrict our estimation to events occurring in the middle phase, specifically $50 \leq t \leq 300$. This heuristic provides robust conclusions: the estimated timing parameters are very close to those found using smaller middle-phase windows, and the global average trend within this window is roughly linear (Fig. A.1a).

For two teams r and b, the outcome of a scoring event (which team receives the point) is given by a biased Bernoulli process, in which the probability that an event increases the score of team i is

$$\Pr(s_i \text{ increases} \,|\, \theta\,) = \begin{cases} c & i = r \\ 1 - c & i = b \ , \end{cases}$$

where $c \in [0, 1]$ represents the competitive advantage (outcome bias) of the r team. In our model system, 99.99% of scoring events yield a single point. Although we do not consider the possibility here, in general, the number of points produced by an event could be drawn from some distribution. Thus, the probability that the competition ends with final scores S_r and S_b is

$$\Pr(S_r, S_b \,|\, c) = c^{S_r}(1 - c)^{S_b} \ , \tag{A.3}$$

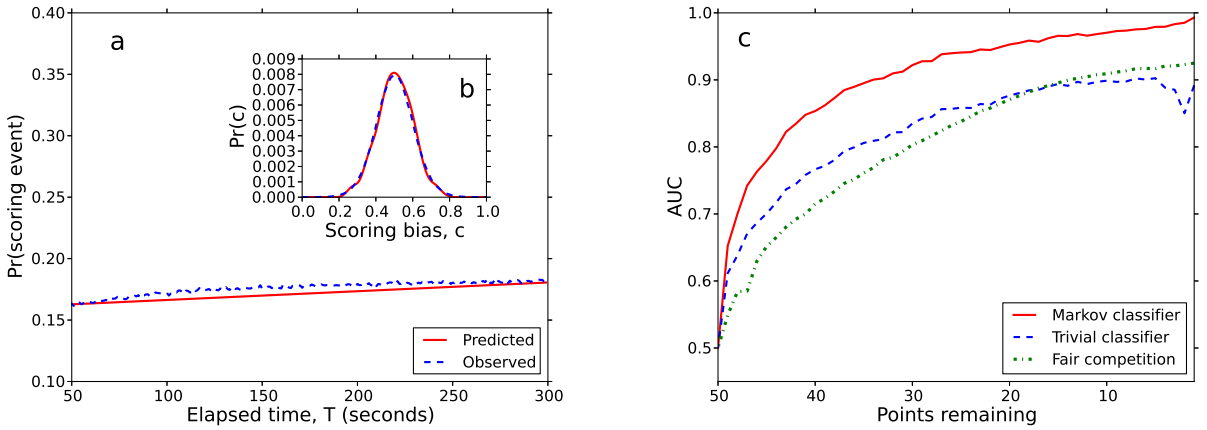

Figure A.1: (a) Global empirical and predicted scoring rates for competitions in *Halo: Reach*, over the window $[50, 300]$ seconds. (A, inset) Global empirical and predicted distribution of competitive advantages (smoothed via a Gaussian kernel). (b) For all competitions, winner predictability (AUC) as a function of team r's points remaining, for three classifiers (see text).

where c denotes the competitive advantage (scoring bias) of team r over team b.

Because team composition varies across competition instances, the competitive advantage of r is modeled as a random variable, drawn from some distribution $\Pr(c)$. The natural choice of the form of this distribution is a symmetric Beta distribution with parameter β, the conjugate prior for the Bernoulli scheme. (We note that the prior distribution must be symmetric about $c = 1/2$ because team labels are arbitrary.) This distributional assumption agrees well with the global empirical distribution of biases c (Fig. A.1a inset).

The posterior probability of observing final scores $\{S_r, S_b\}_k$ in a competition instance k is given by their Bernoulli likelihood, weighted by the probability of c (Eq. (A.3)). Given

parameter		estimate, global	
β	balance	29.50	$\pm$ 0.21
λ_0	base rate	0.1620	$\pm$ 0.0001
α	acceleration	7.00×10^{-5}	$\pm\ 0.05 \times 10^{-5}$

Table A.1: Estimated global scoring tempo and balance parameters, with bootstrap uncertainty estimate.

N such instances, the total posterior probability of the observed final scores is

$$\begin{aligned}
\Pr(\beta \,|\, \{S_r, S_b\}) &= \int_0^1 \left(\prod_{k=1}^{N} \Pr(\{S_r, S_b\}_k \,|\, c) \Pr(c \,|\, \beta) \right) \mathrm{d}c \\
&= \prod_{k=1}^{N} \left(\int_0^1 \frac{c^{S_{r_k}+\alpha-1}(1-c)^{S_{b_k}+\beta-1}}{\mathrm{B}(\beta,\beta)} \,\mathrm{d}c \right) \\
&= \prod_{k=1}^{N} \frac{\mathrm{B}(S_{r_k}+\beta, S_{b_k}+\beta)}{\mathrm{B}(\beta,\beta)} \ , \qquad \text{(A.4)}
\end{aligned}$$

where $\mathrm{B}(a,b)$ is the Beta function.

We estimate the competition balance parameter by numerically maximizing the logarithm of Eq. (A.4) with respect to β,

$$\ln \mathcal{L} = \sum_{k=1}^{N} \ln[\mathrm{B}(S_{r_k}+\beta, S_{b_k}+\beta)] - \ln[\mathrm{B}(\beta,\beta)] \ . \qquad \text{(A.5)}$$

The resulting maximum likelihood estimate $\hat{\beta}$ provides a direct measurement of the overall balance within a set of competition instances: when $\beta \to \infty$, we recover the fair coin $c = 1/2$ expected for ideal competitions.

For a set of competition instances, numerically maximizing Eq. (A.2) with respect to λ_0 and α, and Eq. (A.5) with respect to β, produces maximum likelihood parameter estimates $\hat{\lambda}_0$, $\hat{\alpha}$, and $\hat{\beta}$. Uncertainty in these estimates is then calculated as the standard deviation of the bootstrap distribution [46], where we resample compelte competition instances with replacement. Table A.1 gives the global parameters estimates and uncertainties, when applied to the full set of *Halo: Reach* competitions.

A.3 Predicting Competition Outcomes

For a set of competitions, the predictability of an instance's ultimate winner, after observing only part of the game, provides a second, non-parametric measure non-ideal dynamics. We model scoring as a Markov chain that terminates when a team reaches a score of 50. (In our data, 99% of competitive instances terminate according to this criteria; the remainder from the time limit.)

Suppose an instance has evolved so that teams r and b currently hold scores s_r and s_b. The probability that team r wins the competition is then

$$\begin{aligned}\Pr(r \text{ wins} \,|\, s_r, s_b) = &\Pr(r \text{ wins} \,|\, s_r + 1, s_b) \cdot \hat{c} + \\ &\Pr(r \text{ wins} \,|\, s_r, s_b + 1) \cdot (1 - \hat{c}) \ , \end{aligned} \tag{A.6}$$

where $\hat{c} = s_r/(s_r + s_b)$ is the current maximum likelihood estimate of r's scoring bias within this instance, and the two probability terms capture the probability that r wins if r (or b) wins the next point. (Because a team's score is cumulative, each state in the Markov chain has only two transitions.) Eq. (A.6) is then solved recursively by computing $\hat{c}$ for the current state and working backwards to the instances's current state from the winning states where $s_r = 50$ and $s_b < 50$.

We convert this Markov chain into a classifier by predicting that team r wins if $\Pr(r \text{ wins} \,|\, s_r, s_b) > 0.5$. The probability of correctly choosing the winning team in this case is equivalent to computing the AUC statistic over a set of instances. (AUC is defined as the area under the receiver-operating characteristic (ROC) curve [59], and is mathematically equivalent to the Mann-Whitney U test for distinguishing two classes of items.)

Measuring the AUC as a function of the points remaining provides full information about the way the competition's predictability evolves over time. We convert this information into a point measure by computing, with 40 points remaining for r, the AUC for the Markov classifier, which we then divided by the corresponding AUC for an "ideal" classifier

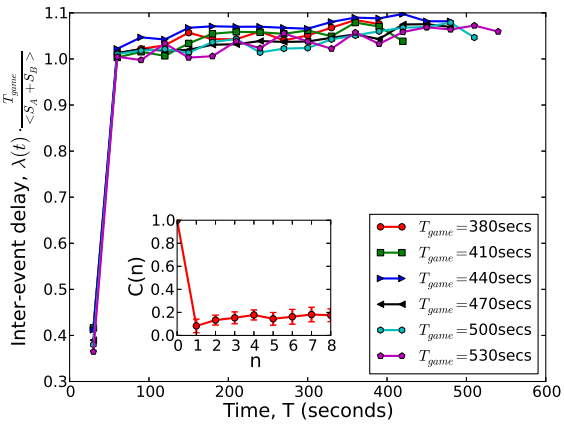

Figure A.2: Average normalized inter-arrival time between scoring events, computed in 30 second intervals, for cohorts of competitions lasting a specific amount of time. (inset) Auto-correlation function $C(n)$ for inter-event times.

(with fixed $c = 1/2$). This provides a direct measure of how much more predictable a real competition's outcome is relative to the ideal model described in the main text.

Using the full data set, Figure A.1B shows the full AUC-over-time curves, for the Markov classifier, the ideal classifier ($c = 1/2$), and for a trivial classifier in which at each moment we predict as the winner the team currently in the lead. Our Markov classifier outperforms the trivial classifier because it captures information about the size of the lead, i.e., it includes information about the bias c in the Bernoulli scoring process, and outperforms the ideal classifier because the competitions' dynamics are non-ideal.

A.4 Test of the Markov Assumption

We now test the accuracy of our Markov assumption in modeling the scoring dynamics of these competitions. If the arrival times of scoring events roughly follow a memoryless Poisson process, there will be little correlation between the sizes of subsequent delays. The correlation function $C(n)$ provides a direct measure of the accuracy of the Markov assump-

tion, and is calculated as

$$C(n) = \frac{\langle T_i T_{i+n} \rangle - \langle T_i \rangle^2}{\langle T_i^2 \rangle - \langle T_i \rangle^2} \ , \tag{A.7}$$

where T_i is the inter-event delay after event i, n is a shift size relative to i, and $\langle . \rangle$ indicates an average over i. A memoryless process matching the Markov assumption in our Bernoulli process will produce $C(n) \approx 0$ for $n > 0$; deviations indicate correlations (or anti-correlations) at the corresponding time scale.

First, a simple rescaling of the observed inter-event delays over the course of competitions of different lengths produces a data collapse (Fig. S2), illustrating relatively little memory in the system. Second, $C(n)$ for our entire sample of competitions (Fig.A.2, inset) shows little correlation (memory) at any time scale. Thus, the Markov assumption seems largely justified.

A.5 Model Goodness-of-Fit

We now test the plausibility of our generative model, i.e., how well it matches the underlying data, by comparing simulated competitions against the empirical data along specific statistical measures. This simulation is parametric and uses the estimated parameters from our generative model to define the corresponding probability distributions in the simulator. A close match between the synthetic scoring dynamics and the empirical data along multiple statistical measure is evidence that our generative model accurately captures the basic features of these competitions.

The simulation framework is given in Algorithm A.5.1. The competition clock is started at $t = 25$ seconds to account for the early-phase delay in the onset of scoring. The bias in the Bernoulli process is then chosen by drawing a value iid from the estimated Beta distribution with parameter $\hat{\beta}$. While neither of the termination criteria have been reached, delays between scoring events are drawn from the estimated linear non-stationary process with parameters $\hat{\lambda}_0$ and $\hat{\alpha}$. Finally, given that a scoring event occurs, with probability c, a single

point is awarded to team r; otherwise, it is awarded to b.

Algorithm A.5.1: COMPETITION SIMULATION()

$t \leftarrow 25$

$s_r \leftarrow s_b \leftarrow 0$

$c \leftarrow$ chooseScoringBias()

while $t < 600$ **and** $s_r < 50$ **and** $s_b < 50$

do
- $T \leftarrow$ interEventDelay()
- **if** $t + T < 600$
  - **then**
    - $\Delta s \leftarrow$ numPoints()
    - updateScores(s_a,s_b, Δ s,c)
    - $t \leftarrow t + T$
  - **else** break

The goodness-of-fit of the model is measured by comparing the simulated and empirical distributions of (i) the final score S, (ii) the final lead size L (at termination), (iii) the number of leader changes m, and (iv) the amount of time t the leading team stays in the lead given a lead of size L. Notably, each of these four quantities is distinct (although related) to the aspects of the data used to estimate the parametric model's structure, and thus they make reasonable checks on the accuracy of the model. Figures A.3a-d show the results of these tests, using 1 million simulated competitions, illustrating very good agreement on all dimensions between simulation and data. Thus, the basic structure of our generative model seems largely justified.

A.6 Additional Results for How Structure Shapes Dynamics

In the main text, we examined four pairs of competition types that each differed on one structural feature: team skill, environmental structure, policies, and resource quality. Figures A.4a-d show the estimated distributions of $\Pr(c)$ (parameterized by $\hat{\beta}$) for these

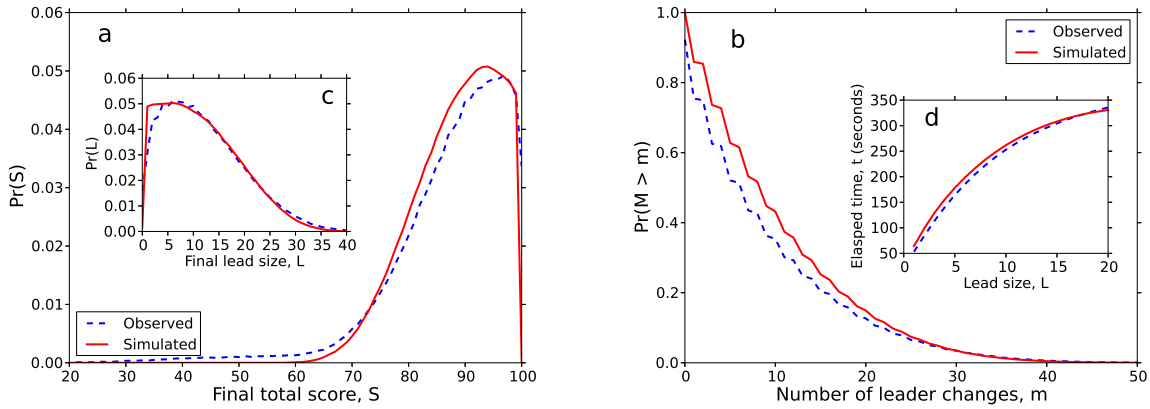

Figure A.3: Comparison of empirical (dashed blue) and simulated (parametric model, red) data for the (a) distribution of final total scores $S = S_r + S_b$, (b) distribution of the number of times the identity of the leading team changes m, (c) distribution of final lead sizes $L = |S_r - S_b|$, and (d) time t elapsed as leader given a lead size of L. The close agreement between data and simulation suggests that our generative model efficiently captures these competitions' dynamics.

four pairs. For each group of instances, the model parameter β was estimated following Section A.2 from the scoring events on the interval $t \in [30, 300]$ seconds of the competition. These times were chosen to exclude biases due to early- and end-phase boundary effects.

Figures A.4e-h show the AUC as a function of points remaining for same competitions, estimated following Section A.3. In each figure, we show for comparison the AUC curve for an ideal competition ($c = 1/2$). The large gap between the Markov classifier's AUC curve and the ideal curve demonstrates that these competitions are substantially more predictable than ideal competitions. This gap is largest early in the competition, where scores are still relatively far from the scoring limit. We also observe modest gaps between the AUC curves for members of each pair, illustrating that structural features do impact the predictability of competition outcomes.

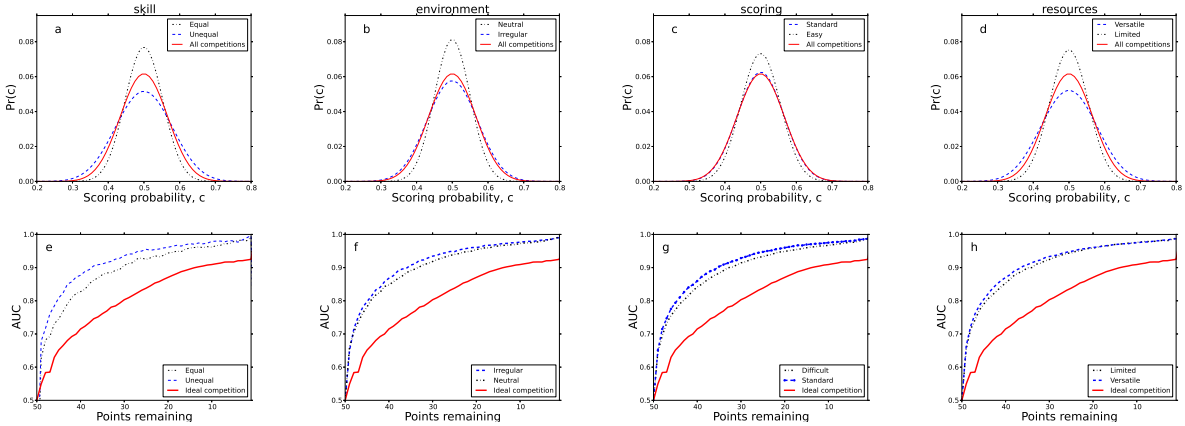

Figure A.4: For the four dimensions discussed in the main text, (A, B, C, D) estimated distribution of scoring biases $\Pr(c)$, and (E, F, G, H) the AUC as a function of points remaining in the competition.

A.7 Additional details of multivariate regression analysis

Here we describe additional details of our investigation of how resources, policy, environment, and skill features explain the variance in the values β, λ_0, α, and ρ observed in our data. To quantify the structure of a competition type $\vec{\eta}$, we defined 35 structural features that characterize the different combinations of environment, resources, policies, and teams. Table A.4 gives the full list of features, with descriptions, classified into four types: resources (R), environment (E), policies (P), and skill (S). Applied to our data yields 125 unique competition types.

For all competition instances with a particular set of features, we estimated the coordinates $(\beta, \lambda_0, \alpha, \rho)$ following Sections A.2 and A.3. Regression models were built on each coordinate independently, and robustness checks were conducted to verify these results (see below). Table A.5 lists the statistically significant ($p \leq 0.1$) features and corresponding coefficients for all four of our models.

For competition balance β, we first used a linear model $\beta = \theta^T \mathbf{x}$, with a design matrix $\mathbf{x}$ composed of the previously defined 125 observations containing 35 features. Fitting this model via least squares produced $r^2 = 0.716$ ($p \ll 0.001$, F-test), but with strongly skewed

residuals. We then fitted the model $\log \beta = \theta^T \mathbf{x}$ to the data, which produced $r^2 = 0.933$ ($p \ll 0.001$, F-test), a marked improvement, and more symmetric residuals. Examining the coefficients, we find that evenly matched teams using medium-to-long-range weapons, competing on large environments without strategic or defensible positions produce more balanced scoring outcomes (larger β).

For the base scoring rate λ_0, a simple linear model yields $r^2 = 0.955$ ($p \ll 0.001$, F-test), indicating that structural features explain almost all the observed variance. The estimated coefficients show that environmental structure features play a dominant role in setting λ_0. In particular, environments that are small, open, and circular correlate best with base scoring rate. In addition to the environment's spatial organization, evenly matched teams also correlate with higher scoring rates. Teams with more experience are likely to be familiar with all terrain options and methods for its exploitation. Environments that are small do not require competitors to spend much time seeking out scoring opportunities (other avatars). Lastly, environments that are open do not provide places to avoid encounters, thus increasing the tempo of competition.

For the acceleration α in the competition tempo, a linear model produces an $r^2 = 0.652$ ($p \ll 0.001$, F-test). We find that few of our features correlate with α, with the exception of long-range weapons and equally-skilled teams, which correlate with smaller α (more ideal competitions). This suggests that in competitions where players are experienced, there is less to learn and thus α is low. This agrees well with the results from λ_0, where more experience leads to a higher base scoring rate.

For the winner predictability ρ, a linear model produces an $r^2 = 0.885$ ($p \ll 0.001$, F-test). Notably, features related to neutral environments and equally-skilled teams correlated with less predictable (more ideal) outcomes. As expected from the correlation between β and ρ (Table A.2), features that correlated with greater β typically also correlate with lower ρ.

Finally, we expected changes in policy to have an impact on scoring balance and tempo

of events. However, we find that policy type features do not by themselves play a role in controlling these dynamics, once we control for other variables like skill, environmental structure and resources. Specifically, we find that the policy feature coefficients are insignificant in all of our models ($p > 0.1$) and thus we excluded from the results of our best-subset selection.

A.7.1 Tests of model robustness

To test the robustness of our results against spurious correlation, due to the high-dimensionality of our data, we conducted three additional analyses.

First, we consider colinearity among the dependent variables. Table A.2 lists the pairwise coefficients of variation r^2, showing a high degree of correlation between ρ and $\log \beta$, modest correlation between $\log \beta$ and λ_0, but little else. To test whether these correlations impact our results, we conducted a MANOVA on a multiple multivariate regression model (Table A.6). The results show that the same set of features reported in Table A.5 are significant, suggesting that our original results are robust.

Second, we perform a stepwise AIC feature selection procedure to choose the best subset of features under mild regularization. With the exception of α, the results shown in Tables A.7, A.8, and A.9 indicate that the selected features and their weights presented in the original regression analysis are robust. The best-subset selection for α produces a larger list of significant features than in the original model, but a slightly lower r^2. The most significant negative feature, long range resources, is robust to this procedure while equally skilled teams and other resource features are not.

Finally, we perform a randomization test by randomly permuting the dependent variables across the associated features and repeating the original multivariate regression. This randomization destroys any natural correlation between the features and the dependent variable. Table A.3 shows the resulting coefficients of variation, none of which are statistically significant. These results further support the robustness of our original results.

	$\log\beta$	λ_0	α	ρ
$\log\beta$	–	0.356	0.053	0.776
λ_0	0.356	–	0.003	0.398
α	0.053	0.003	–	–
ρ	0.776	0.398	–	–

Table A.2: Coefficients of variation r^2 for pairs of dependent variables. Cells containing no data are either irrelevant or statistically insignificant ($p > 0.1$).

parameter	r^2	p-value
$\log\beta$	0.08	0.98
λ_0	0.12	0.84
α	0.12	0.8
ρ	0.08	0.98

Table A.3: Regression results after randomly permuting the vectors of 35 independent variables and tuple of 5 scoring dynamics parameters, $(\log\beta, \lambda_0, \alpha, \rho)$.

A.8 Player preference and competition balance

When competitions are predictable they become less enjoyable. In professional sports this manifests itself as fans leaving a stadium well before the end of a game when one team is winning by such a large amount that there is little chance that the trailing team will make a comeback.

In our model system, the same decision can occur for players themselves, who can effectively walk off the field by voluntarily exiting the competition early. For each of the competition types in our sample we calculated the competition dropout rate as

$$\omega = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{\text{at least 1 player quits early}\}, \tag{A.8}$$

where N is the number of instances of the given type.

From the first 25 million games, we extracted a total of 4.1 million competitive type games that did not contain corrupt data. From these 4.1 million games we selected only those where at least one player left the game early. Using the remaining 1.9 million games we then

tested for a correlation between the dropout rate ω and the overall balance β. If players prefer more balanced competitions, as β increases (more ideal competitions), the dropout rate should decrease. A simple linear regression yields the equation $\ln \omega = 1.593 - 1.371 \ln \beta$ ($r^2 = 0.43$, $p \ll 0.001$, t-test). These results corroborate our hypothesis, illustrating that the more predictable the scoring dynamics of a competition (small β), the more likely at least one player will exit early. Quantitatively, this relationship predicts that increasing competition balance β by a factor of 1.66 correlates with reducing the early exit probability ω by a factor of 2.

As a caveat, we note that there are several involuntary reasons a player may exit early, e.g., network issues, power loss, system error, being "booted" for excessive friendly fire, and several voluntary reasons unrelated to player engagement, e.g., to join friends in another game, to change competition types, etc. Most of these variables are inaccessible to us for analysis; however, we cannot conceive of a mechanistic relationship between most of these reasons and the scoring balance of a competition. Additional investigation may further illuminate the precise mechanism by which increase in β produce decreased exit rates.

Table A.5: Ordered multivariate regression coefficients for standard ("slayer") competitions for each estimated model parameter $\log \beta$, λ_0, α, and ρ.

parameter	feature	θ	std. error	t value	$\Pr(>\|t\|)$	r^2
$\log \beta$	E5	1.849	0.320	5.764	$\ll 0.001$	0.933
	E1	1.391	0.371	3.745	$\ll 0.001$	
	E11	1.123	0.141	7.920	$\ll 0.001$	
	S1	0.822	0.034	23.828	$\ll 0.001$	
	E3	0.570	0.256	2.224	0.028	
	E9	0.481	0.076	6.265	$\ll 0.001$	

Continued on next page

Table A.5 – *Continued from previous page*

parameter	feature	θ	std. error	t value	$\Pr(>\|t\|)$	r^2
$\log \beta$	E5	1.849	0.320	5.764	$\ll 0.001$	0.933
	R10	−0.354	0.134	−2.642	0.009	
	R8	−0.495	0.215	−2.303	0.023	
	R15	−0.580	0.233	−2.488	0.014	
	E6	−0.813	0.150	−5.414	$\ll 0.001$	
	E2	−1.861	0.252	−7.375	$\ll 0.001$	
	E7	−2.126	0.224	−9.467	$\ll 0.001$	
λ_0	E5	0.082	0.008	9.966	$\ll 0.001$	0.955
	E11	0.059	0.003	16.344	$\ll 0.001$	
	E1	0.045	0.009	4.774	$\ll 0.001$	
	E3	0.029	0.006	4.437	$\ll 0.001$	
	E9	0.023	0.001	12.028	$\ll 0.001$	
	R10	0.008	0.003	2.478	0.014	
	S1	0.005	0.001	6.010	$\ll 0.001$	
	E4	−0.009	0.004	−2.374	0.019	
	R8	−0.011	0.005	−1.995	0.048	
	R13	−0.011	0.004	−2.266	0.025	
	E6	−0.011	0.003	−2.845	0.005	
	R2	−0.015	0.008	−1.873	0.063	
	R1	−0.021	0.008	−2.680	0.008	
	R4	−0.030	0.008	−3.797	$\ll 0.001$	
	R15	−0.032	0.006	−5.444	$\ll 0.001$	
	E2	−0.081	0.006	−12.448	$\ll 0.001$	
	E7	−0.081	0.005	−13.991	$\ll 0.001$	

Continued on next page

Table A.5 – *Continued from previous page*

parameter	feature	θ	std. error	t value	$\Pr(>\|t\|)$	r^2
α	R12	-1.9×10^{-5}	8.1×10^{-6}	-2.449	0.016	0.652
	S1	-2.9×10^{-6}	1.7×10^{-6}	-1.692	0.093	
ρ	E7	0.138	0.022	6.295	$\ll 0.001$	0.885
	E2	0.123	0.024	4.989	$\ll 0.001$	
	R4	0.070	0.030	2.299	0.023	
	E6	0.061	0.014	4.175	$\ll 0.001$	
	R1	0.053	0.030	1.734	0.085	
	R15	0.046	0.022	2.030	0.044	
	R8	0.040	0.021	1.937	0.055	
	E4	0.031	0.015	2.018	0.046	
	R3	0.029	0.015	1.852	0.066	
	R14	-0.030	0.012	-2.366	0.019	
	E9	-0.036	0.007	-4.775	$\ll 0.001$	
	S1	-0.055	0.003	-16.413	$\ll 0.001$	
	E11	-0.089	0.013	-6.410	$\ll 0.001$	
	E5	-0.095	0.031	-3.020	0.003	

	feature	code	domain	description
resources	loadout_1	R1	$\{0,1\}$	short range and medium range
	loadout_2	R2	$\{0,1\}$	low quality resources
	loadout_3	R3	$\{0,1\}$	long range and grenades
	loadout_4	R4	$\{0,1\}$	short and long range
	loadout_5	R5	$\{0,1\}$	medium range
	vehicles_revenant	R6	$\{0,1\}$	lightly armored vehicle
	vehicles_scorpion	R7	$\{0,1\}$	heavy tank vehicle
	vehicles_mongoose	R8	$\{0,1\}$	unarmored vehicle
	vehicles_ghost	R9	$\{0,1\}$	rapid attack vehicle
	weapons_short	R10	$\{0,1\}$	short range
	weapons_medium	R11	$\{0,1\}$	medium range
	weapons_long	R12	$\{0,1\}$	long range
	weapons_grenades	R13	$\{0,1\}$	grenade type
	weapons_rocket	R14	$\{0,1\}$	rocket launcher
	weapons_unsc	R15	$\{0,1\}$	high-quality only resources
	weapons_covenent	R16	$\{0,1\}$	low-quality only resources
	weapons_both	R17	$\{0,1\}$	high- and low-quality resources
skill	TrueSkill matchmaking	S1	$\{0,1\}$	equally skilled teams
	team size	S2	$\{0,1\}$	4- or 5-person teams
environmental structure	map_open	E1	$\{0,1\}$	open terrain
	map_vertical	E2	$\{0,1\}$	vertical environment
	map_circular	E3	$\{0,1\}$	circular terrain
	map_varied	E4	$\{0,1\}$	no clear organizing principle
	map_corridors	E5	$\{0,1\}$	indoor terrain
	map_bases	E6	$\{0,1\}$	defensible positions
	map_towers	E7	$\{0,1\}$	high ground
	map_transporters	E8	$\{0,1\}$	teleporters, jump pads and vents
	map_outdoor	E9	$\{0,1\}$	outdoor terrain
	map_size_small	E10	$\{0,1\}$	small or medium sized map
	map_size_large	E11	$\{0,1\}$	large arena
	map_size_perim	E12	$\mathbb{R}^+$	perimeter of map, seconds to run
policies	rules_noradar	P1	$\{0,1\}$	HUD radar is off
	rules_noshields	P2	$\{0,1\}$	shield is off
	rules_headshot	P3	$\{0,1\}$	headshot required for kill (SWAT rules)
	rules_snipers	P4	$\{0,1\}$	sniper fight

Table A.4: Competition features, abbreviations and verbal descriptions, grouped in four categories: resources (R), skill (S), environmental structure (E), and policy (P).

feature	df	Wilks	approx. F	num. df	den. df	$\Pr(> F)$
R1	1	0.533	21.617	4	99	$\ll 0.001$
R2	1	0.339	48.147	4	99	$\ll 0.001$
R3	1	0.352	45.541	4	99	$\ll 0.001$
R4	1	0.716	9.802	4	99	$\ll 0.001$
R8	1	0.167	123.322	4	99	$\ll 0.001$
R10	1	0.302	57.109	4	99	$\ll 0.001$
R11	1	0.418	34.459	4	99	$\ll 0.001$
R12	1	0.383	39.799	4	99	$\ll 0.001$
R13	1	0.817	5.536	4	99	$\ll 0.001$
S1	1	0.112	194.402	4	99	$\ll 0.001$
R15	1	0.224	85.703	4	99	$\ll 0.001$
E1	1	0.455	29.610	4	99	$\ll 0.001$
E2	1	0.358	44.342	4	99	$\ll 0.001$
E3	1	0.606	16.076	4	99	$\ll 0.001$
E4	1	0.811	5.742	4	99	$\ll 0.001$
E5	1	0.246	75.711	4	99	$\ll 0.001$
E6	1	0.399	37.133	4	99	$\ll 0.001$
E7	1	0.842	4.623	4	99	0.001
E9	1	0.401	36.896	4	99	$\ll 0.001$
E11	1	0.239	78.378	4	99	$\ll 0.001$

Table A.6: MANOVA results of multiple multivariate regression model, providing a robustness check on the results given in Table A.5.

parameter	feature	θ	std. error	t value	$\Pr(>\|t\|)$	r^2
$\log\beta$	E5	1.803	0.229	7.867	$\ll 0.001$	0.933
	E1	1.320	0.228	5.779	$\ll 0.001$	
	E11	1.126	0.124	9.029	$\ll 0.001$	
	S1	0.822	0.034	24.153	$\ll 0.001$	
	E3	0.480	0.122	3.919	$\ll 0.001$	
	E9	0.479	0.069	6.888	$\ll 0.001$	
	R13	0.154	0.069	2.243	0.027	
	R14	0.119	0.074	1.598	0.113	
	R1	-0.322	0.054	-5.952	$\ll 0.001$	
	R3	-0.232	0.092	-2.505	0.013	
	R12	-0.310	0.110	-2.822	0.005	
	R10	-0.367	0.113	-3.232	0.001	
	R8	-0.472	0.181	-2.596	0.01	
	R4	-0.504	0.062	-8.081	$\ll 0.001$	
	R15	-0.644	0.092	-6.931	$\ll 0.001$	
	E6	-0.827	0.130	-6.353	$\ll 0.001$	
	E2	-1.860	0.207	-8.957	$\ll 0.001$	
	E7	-2.093	0.193	-10.840	$\ll 0.001$	
λ_0	E5	0.084	0.006	13.770	$\ll 0.001$	0.954
	E11	0.061	0.002	20.759	$\ll 0.001$	
	E3	0.029	0.003	8.648	$\ll 0.001$	
	E9	0.024	0.001	12.383	$\ll 0.001$	
	R10	0.008	0.003	2.794	0.006	
	R3	0.005	0.002	2.080	0.039	
	S1	0.005	0.001	6.085	$\ll 0.001$	
	E1	0.048	0.005	8.880	$\ll 0.001$	
	R13	-0.009	0.002	-3.979	$\ll 0.001$	
	E4	-0.008	0.002	-3.178	0.001	
	R8	-0.011	0.004	-2.467	0.015	
	E6	-0.012	0.003	-3.860	$\ll 0.001$	
	R2	-0.015	0.005	-2.939	0.004	
	R1	-0.022	0.005	-4.191	$\ll 0.001$	
	R4	-0.031	0.005	-5.852	$\ll 0.001$	
	R15	-0.034	0.004	-8.469	$\ll 0.001$	
	E7	-0.080	0.004	-16.695	$\ll 0.001$	
	E2	-0.081	0.005	-14.457	$\ll 0.001$	

Table A.7: Ordered multivariate regression model coefficients for all standard ("slayer") competitions regressed onto $\log\beta$, λ_0, selected via stepwise AIC, providing a second check on the robustness of the results in Table A.5.

parameter	feature	θ	std. error	t value	$\Pr(>\|t\|)$	r^2
ρ	E7	0.124	0.010	11.934	$\ll 0.001$	0.882
	E2	0.111	0.011	9.943	$\ll 0.001$	
	R4	0.067	0.010	6.444	$\ll 0.001$	
	E6	0.052	0.005	8.998	$\ll 0.001$	
	R1	0.049	0.010	4.958	$\ll 0.001$	
	R8	0.046	0.016	2.779	0.006	
	R15	0.045	0.006	7.335	$\ll 0.001$	
	E4	0.039	0.007	5.456	$\ll 0.001$	
	R2	0.037	0.010	3.533	$\ll 0.001$	
	R3	0.027	0.008	3.420	$\ll 0.001$	
	E9	-0.034	0.006	-4.912	$\ll 0.001$	
	R14	-0.036	0.006	-5.971	$\ll 0.001$	
	S1	-0.055	0.003	-16.763	$\ll 0.001$	
	E5	-0.076	0.010	-7.429	$\ll 0.001$	
	E11	-0.081	0.006	-12.389	$\ll 0.001$	

Table A.8: Ordered multivariate regression model coefficients for all standard ("slayer") competitions regressed onto ρ selected via stepwise AIC, providing a second check on the robustness of the results in Table A.5.

parameter	feature	θ ($\times 10^{-5}$)	std. error ($\times 10^{-6}$)	t value	$\Pr(>\|t\|)$	r^2
α	R3	1.570	2.583	6.077	$\ll 0.001$	0.637
	R11	1.446	3.328	4.345	$\ll 0.001$	
	R2	1.432	2.965	4.832	$\ll 0.001$	
	E5	1.105	2.114	5.226	$\ll 0.001$	
	E3	0.454	2.368	1.918	0.057	
	S1	-0.294	1.689	-1.746	0.083	
	R1	-0.470	2.529	-1.859	0.065	
	R15	-1.591	2.583	-6.157	$\ll 0.001$	
	R8	-1.868	7.159	-2.609	0.010	
	R12	-2.551	2.538	-10.053	$\ll 0.001$	

Table A.9: Ordered multivariate regression model coefficients for all standard ("slayer") competitions regressed onto α selected via stepwise AIC, providing a second check on the robustness of the results in Table A.5.

Appendix B

Scoring dynamics across professional team sports: tempo, balance and predictability

Supporting Information

Here we provide brief summaries of the game mechanics for each of the sports represented in our data set.

Professional and college football

Both professional and college football games last 60 minutes, divided into four equal length "quarters." Each of the two teams field 11 players, the identities of which are usually changed depending on whether a team has possession of the ball (offense) or not (defense). The field is a flat grass or turf surface 360 feet long and 160 feet wide. On either end of the field are two "end zones," measuring 30 feet in length, one for each team.

The offense is given a series of four attempts ("downs") to move the football 10 yards downfield from the last valid ball position, each of which must occur within 40 seconds of the last attempt's end. Failure to move the ball the required distance results in the other team gaining possession. Points are scored by the team in possession when it moves the ball into the defensive team's "end zone," (a touchdown, 6 points) or passes the ball through the defensive team's field goal (a field goal, 3 points). Scoring a touchdown provides the scoring team the opportunity for additional points, either through what would normally be a touchdown (2 points) or a field goal (1 point). Each team has three timeouts to use during gameplay, which are often used strategically near the end of the game. When time runs

out, the team with the most points is declared the winner. For a complete description of professional and college level rules, see [91] and [90] respectively.

Professional hockey

A professional hockey game lasts 60 minutes, divided into three equal length "periods." A game is played between two teams, each composed of six players (five skaters and one goalkeeper), whose identities change periodically throughout the game. Teams compete on an ice rink, 200 feet long and 85 feet wide. On either end of the rink are two nets, 6 feet wide and 4 feet high, one for each team.

Players on the team controlling the puck work together to move it into the opposing team's net through a combination of strategic passes and shots. If the team is successful, a goal is scored and the team is awarded 1 point. The game plays continuously except after stoppages, which occur at minutes 6, 10, and 14, penalties, or goals. Each team has a single 30 second timeout that can be used at any point in the game. Teams use their timeouts to substitute players, adjust strategy and to provide the team with brief moments of rest during crucial periods of play. When time runs out, the team with the most points is deemed the winner. For a complete description of rules, see [92].

Professional basketball

A professional basketball game lasts 48 minutes, divided into four equal length "quarters." Each of the two teams field five players, whose identities change throughout the game. The court is a flat wooden surface, 94 feet long and 50 feet wide. On either side of the court are two circular rims, known as baskets, measuring 18 inches in diameter, positioned 10 feet above the court surface, one for each team.

A team in possession of the basketball has a total of 24 seconds to make a shot that it either hits the opposing team's rim or goes through it. If time expires before the team attempts a shot, the opposing team gains possession of the basketball. Depending on game

state and a player's court location, a successful shot (one that goes through the opposing team's rim) can be worth 1, 2, or 3 points. After scoring, the scoring team relinquishes possession of the basketball to the opposing team. Game play continues according to this procedure, except when the ball goes out of bounds or a foul is committed. Each team is awarded a single 20 second timeout per game half. Each team is also entitled to 6 more timeouts that may be used at anytime throughout the game, with the following restrictions: no more than 3 timeouts may be used during the final quarter and no more than 2 timeouts may be used within the final 2 minutes of play. These timeouts are used strategically to substitute players, control the speed of play, and facilitate the coordination and planning of complex plays. When time expires, the team that has accumulated the most points is deemed the winner. For a full description of game rules, see [89].

B.1 Points per scoring event

Table B.1 shows the distribution of points per scoring event, for each sport. Events in the NHL only generate a single point. Although events in the NBA generate 1, 2 or 3 points, the large majority of events (74%) are worth 2 points, with the remaining events divided between 1- and 3-point shots.

Similarly, scoring events in both CFB and NFL games generally produce 7 points (touchdown with extra point). Games in CFB games from those in NFL in producing many more field goals (3 points) and many fewer touchdowns with no extra point (6 points), which are the next most common events in both. The remaining point values are relatively uncommon: 8 points for touchdowns plus a 2 point conversion play, and 2 points for a safety, which occurs in three scenarios: (i) when a ball carrier is tackled in his team's own end zone; (ii) when the ball is deemed dead by referees in the end zone, or (iii) when the offensive team commits a foul play in its own end zone. Two point conversions occur when the scoring team elects to successfully pass or run the ball into the end zone instead of kicking the ball through the field goal after a touchdown.

Figure B.1 shows the fraction of total points in each game that are won by a team, which agrees very closely with the fraction of total scoring events, from Figure 6.4. This agreement indicates that only very rarely does the value of the points associated with events ultimately determine the outcome of a game. Instead, the chief determinant is simply number of events. In NHL games, this must be true as every event is worth the same number of points. A slightly deviation around 1/2 for NFL games, but not CFB games, indicates that very occasionally point values do matter.

point value	NFL	CFB	NHL	NBA
1	-	-	**1.0000**	0.0941
2	0.0083	0.0113	-	**0.7373**
3	0.3055	0.1702	-	0.1647
4	-	-	-	0.0029
5	-	-	-	0.0009
6	0.0308	0.0708	-	0.0001
7	**0.6222**	**0.7058**	-	-
8	0.0332	0.0419	-	-
any	1.0000	1.0000	1.0000	1.0000

Table B.1: Empirical distribution of all regulation scoring events over point values, by sport, with the modal value highlighted.

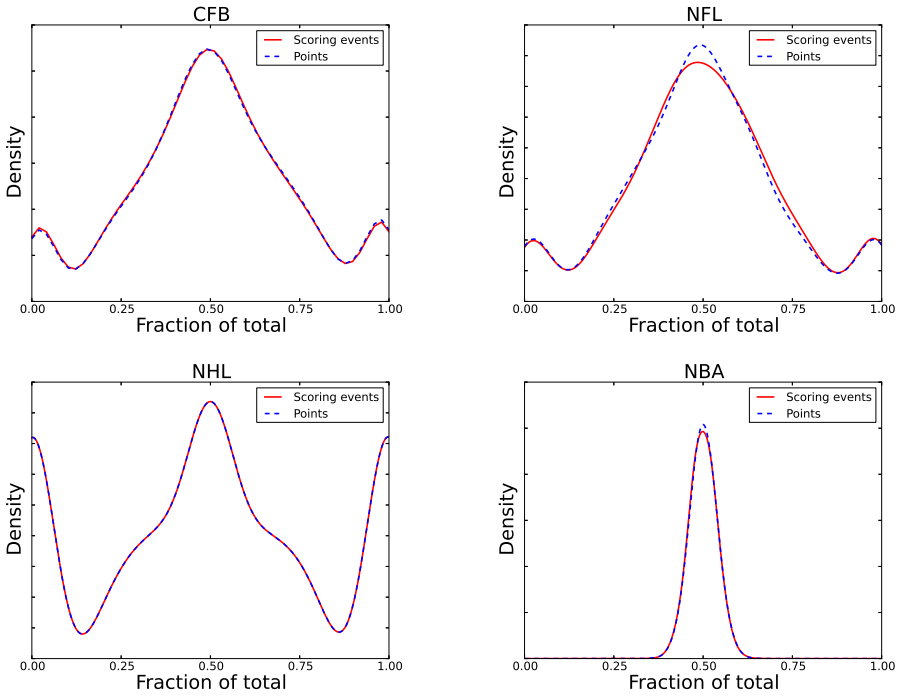

Figure B.1: Smoothed distributions for the empirical fraction of total points won by a team (solid line), for each sport, plus the empirical fraction of total scoring events (dashed line; from Figure 6.4). The very close agreement indicates that only very rarely does the point-value of scoring events—instead of simply their number—determine the outcome of a game.

Appendix C

Occupation networks: career path dynamics and income traps

Supporting Information

Figure C.1 displays word clouds for each of the 5 largest communities studied here. The large words in each cloud are those that appear most frequently in the data.

Figure C.1: Word clouds for the sets of occupation titles contained in each of the 5 largest communities in the occupation network. The larger the word the more frequent it appears in titles. Words that appear in multiple clouds suggest that occupation titles containing them act as bridges, inter-connecting each community.